Lecture 02. Classification (Chapter 4)

Ping Yu

HKU Business School
The University of Hong Kong

An Overview of Classification

(Section 4.1)

Classification

- Qualitative variables take values in an unordered set $\mathscr{C}$, such as:

  eye color $\in$ {brown, blue, green}
  email $\in$ {spam, ham} (ham=good email)
  marital status $\in$ {married, single,divorced, widowed}
  digit $\in$ {0, 1, $\cdots$,9}

- Given a feature vector X and a qualitative response Y taking values in the set $\mathscr{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e., $C(X) \in \mathscr{C}$.
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathscr{C}$.

  - For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

  - Such a classifier behaves like regression methods because $\Pr(Y = 1|X = x) = E[I(Y = 1)|X = x]$, just like $E[Y|X = x]$.

[Example] Credit Card Default

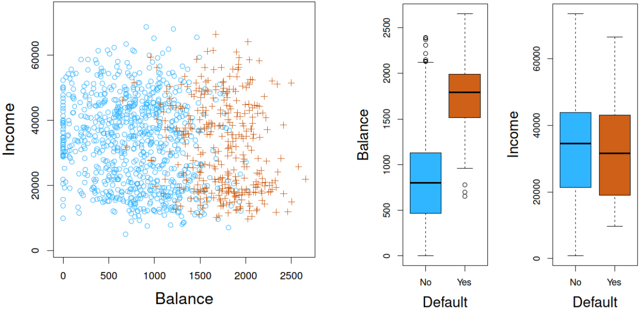


FIGURE 4.1. *The* `Default` *data set.* Left: *The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.* Center: *Boxplots of* `balance` *as a function of* `default` *status.* Right: *Boxplots of* `income` *as a function of* `default` *status.*

- $Y \in \{\text{No}, \text{Yes}\}$ is binary, $X \in \mathbb{R}_+^2$; $\bar{y} = 3\%$, so only part of the $y_i = 0$ data are shown.

Some Popular Classifiers Covered in This Lecture

- logistic regression
- linear discriminant analysis (LDA)
- quadratic discriminant analysis (QDA)
- naive Bayes
- K-nearest neighbors (KNN)

- Many other computer-intensive classification methods would be covered in later lectures, e.g., support vector machines build structured models for $C(x)$.

Why Not Linear Regression?

(Section 4.2)

Can We Use Linear Regression?

- Suppose for the Default classification task that we code

$$Y = \begin{cases} 0, & \text{if No}, \\ 1, & \text{if Yes}. \end{cases}$$

- Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?
- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to LDA which we discuss later. [Exercise*]
- Since in the population $E\left[Y|X=x\right] = \Pr(Y=1|X=x)$, we might think that regression is perfect for this task.
- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate. [Figure here]

Linear versus Logistic Regression

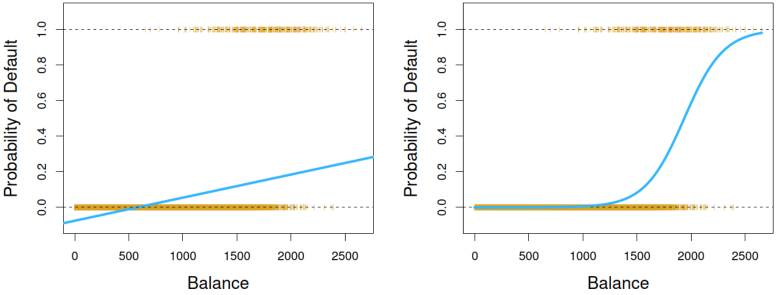


FIGURE 4.2. *Classification using the* Default *data.* Left: *Estimated probability of* default *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* default*(*No *or* Yes*).* Right: *Predicted probabilities of* default *using logistic regression. All probabilities lie between* 0 *and* 1.

Y Takes More Than 2 Values

- Beside $\hat{Y} \notin [0,1]$ such that $\hat{Y}$ cannot be give a probability interpretation, a regression method cannot accommodate a qualitative response with more than two classes.
- First, if Y is not ordered, it is hard to assign meaningful values to Y; e.g., for eye color, if we assign 1, 2 and 3 to brown, blue, and green, why not 3, 2 and 1, and why not 1, 3, 7?
  - For different assignments, the prediction results are different.
  - How about $Y \in \{0,1\}$? Which outcome of Y is assigned 1 does not matter for prediction! [Exercise]
- Second, if Y is ordered such as $Y \in \{$bad, neutral, good$\}$, how do we know the difference between bad and neutral is the same as that between neutral and good if we assign $\{-1,0,1\}$ to $\{$bad, neutral, good$\}$?
- Linear regression is not appropriate here. Multiclass Logistic Regression or Discriminant Analysis are more appropriate.

Logistic Regression

(Section 4.3)

History of Logistic Regression

Joseph Berkson (1899-1982, Mayo Clinic)

- Berkson introduced the logit model (and coined the term "logit") in 1944 as a convenient computational approximation to the probit (“probability unit”) model.

The Logistic Model

- Logistic regression does not model Y directly, but models the probability that Y belongs to a particular category (so the values assigned to Y do not matter).
- Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using balance to predict default. Logistic regression uses the logistic function[1]

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}, \tag{1}$$

where $e \approx 2.71828$ is the Euler's number.

- It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1 [see Figure 4.2; how β_0 and β_1 affect the shape of $p(X)$?], and $\frac{\partial p(X)}{\partial X}$ is not constant [Exercise].
- Solving out $\beta_0 + \beta_1 X$ in (1) gives

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \in \mathbb{R}.$$

- This monotone increasing transformation of $p(X)$ is called the log odds [2]or logit transformation of $p(X)$, so rather than modeling $p(X)$ as a linear function of X, we model $\text{logit}(p(X))$ linearly here.

[1] The *logistic function* $\frac{e^X}{1+e^X}$ was introduced by Pierre François Verhulst in 1938 as a modified exponential growth model. It is speculated that he used the term *logistic* as a contrast to *logarithmic*.

[2] by log we mean natural log: ln.

Maximum Likelihood

- Maximum likelihood is another basic principle of econometrics besides least squares; it guesses the most likely reason (i.e.,the parameter value) for a phenomenon (i.e., the observed data).
- Here, we use maximum likelihood to estimate the unknown parameters (β_0, β_1).
- The likelihood function is

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}.$$

- This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.
- Essentially, we try to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that plugging these estimates into the model for $p(X)$, given in (1), yields a number close to 1 for all individuals who defaulted, and a number close to 0 for all individuals who did not. [Exercise, see also Figure 4.2]
- Most statistical packages can fit linear logistic regression models by maximum likelihood. In **R** we use the glm function. [$\widehat{\beta}_1 = \frac{\partial \text{log odds}}{\partial X}$, and sign$\left(\widehat{\beta}_1\right)$ =sign$\left(\frac{\partial \hat{p}(X)}{\partial X}\right)$]

	Coefficient	Std. error	z-statistic	p-value
Intercept	-10.6513	0.3612	-29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

Making Predictions

- What is our estimated probability of default for someone with a balance of \$1000?
- Since

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}},$$

  we have

$$\hat{p}(1000) = \frac{e^{-10.6513 + 0.0055 \times 1000}}{1 + e^{-10.6513 + 0.0055 \times 1000}} = 0.006;$$

  similarly, $\hat{p}(2000) = 0.586$. [see Figure 4.2]
- Let's do it again, using the dummy variable student as the predictor.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\text{default} = \text{Yes} | \text{student} = \text{Yes}) = \frac{e^{-3.5041 + 0.4049 \times 1}}{1 + e^{-3.5041 + 0.4049 \times 1}} = 0.0431,$$

$$\widehat{\Pr}(\text{default} = \text{Yes} | \text{student} = \text{No}) = \frac{e^{-3.5041 + 0.4049 \times 0}}{1 + e^{-3.5041 + 0.4049 \times 0}} = 0.0292.$$

Multiple Logistic Regression

- Like the extension from simple to multiple linear regression, we generalize

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p=:\beta_0+\beta^TX\Rightarrow p(X)=\frac{e^{\beta_0+\beta^TX}}{1+e^{\beta_0+\beta^TX}},$$

where $X=(X_1,\cdots,X_p)^T$ are p predictors, $\beta=\left(\beta_1,\cdots,\beta_p\right)^T$ collects the slopes.

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

- Why is coefficient for student negative, while it was positive before?
- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students. [see right panel of Figure 4.3]
- But for each fixed level of balance, students default less than non-students [see left panel of Figure 4.3]; the effect of student is confounded by balance.
- Multiple logistic regression can tease this out.

Confounding

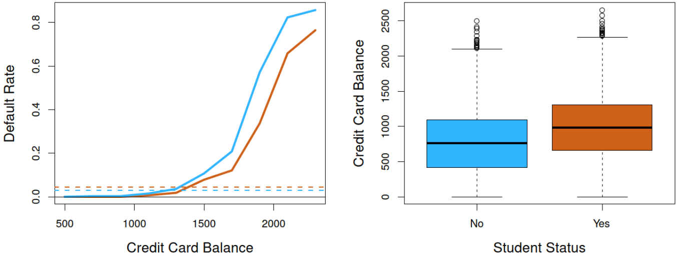


FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

- Should the credit card company offer credit to a student? Depending on what variables are controlled for; ideally, the confounding variables (i.e., the variables that can explain Y and meanwhile correlate with the interested X_j) should be included in the logistic regression.

Multinomial/Multiclass Logistic Regression

- When there are $K > 2$ classes, we choose one class, say the Kth class, as the baseline, and model

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \tag{2}$$

for $k = 1, \cdots, K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}},$$

where note that $\beta_k := \left(\beta_{k0}, \beta_{k1}, \cdots, \beta_{kp}\right)^T$ depends on k.

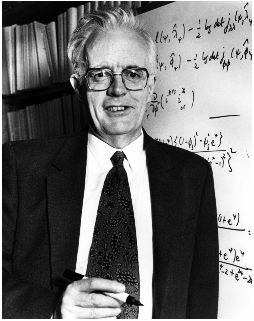
- When $K = 2$, we choose the class with $Y = 0$ as the baseline.

- It is not hard to see that

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p.$$

- When $K > 2$, for different choices of baseline class, the predictions, log odds and other key model outputs are the same although $\hat{\beta}$ may be different.

History of Multinomial Logistic Regression



David R. Cox
(1924-2022, Oxford)

Henri Theil
(1924-2000,
Chicago and Florida)

Daniel L. McFadden
(1937-, UC-Berkeley,
NP2000)

- The multinomial logit model was introduced independently in Cox and Theil.
- McFadden linked the multinomial logit model to the theory of discrete choice, which gave a theoretical foundation for logistic regression.

Softmax Coding

- **R** package glmnet employs the following symmetric form, called the softmax coding,

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}, \quad (3)$$

  where there is a linear function for each class.
  - In (2), the baseline β_K is set as **0**.
- There is some redundancy in the parametrization of softmax coding.
- Some cancellation is possible, and only $K-1$ linear functions are needed as in 2-class logistic regression.
- It is not hard to see that the log odds ratio between the kth and k'th classes equals

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0}-\beta_{k'0}) + (\beta_{k1}-\beta_{k'1})x_1 + \cdots + \left(\beta_{kp}-\beta_{k'p}\right)x_p,$$

  so the coefficients β_k in (2) are actually the difference of β_k and β_K in (3), which is why β_K is normalized as **0**.

Generative Models for Classification

(Section 4.4)

History of Discriminant Analysis

Ronald A. Fisher (1890-1962), UCL

Generative Model vs Discriminative Model

- Discriminative models draw boundaries in the data space, while generative models try to model how data is placed throughout the space. [figure here]
  - A generative model could generate new photos of animals that look like real animals, while a discriminative model could tell a dog from a cat. (GPT)
- A generative model focuses on explaining how the data was generated, while a discriminative model focuses on predicting the labels of the data.
- Mathematically, given a set of data instances X and a set of labels Y, generative models capture the joint probability $\Pr(X, Y)$, or just $\Pr(X)$ if there are no labels, while discriminative models capture the conditional probability $\Pr(Y|X)$.
- (**) A generative model is helpful to deal with missing input values, outliers and unlabelled data points which requires access to $p(x)$ and $p(x) = \sum_y p(y) p(x|y)$ in a generative model.

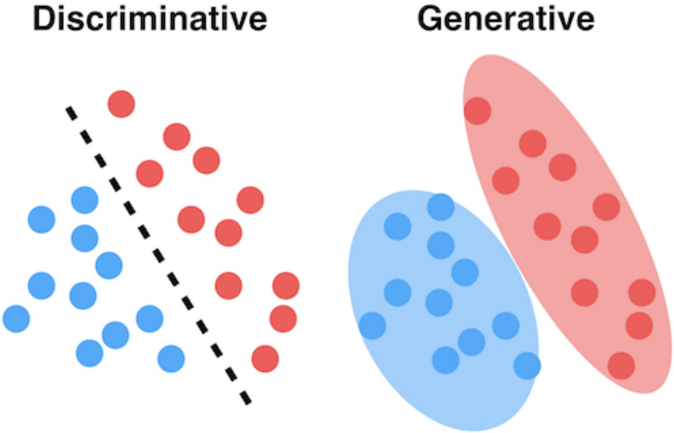


Figure: Difference Between Discriminative Model and Generative Model

Discriminant Analysis

- Here the approach is to model the distribution of X in each of the classes separately, and then use Bayes' theorem to flip things around and obtain $\Pr(Y|X)$. [Figure here]
- When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.
- However, this approach is quite general, and other distributions can be used as well. We will focus on normal distributions.

Why Discriminant Analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable [Why?]. LDA does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the LDA is again more stable than logistic regression.
- LDA is popular when $K > 2$, because it also provides low-dimensional views of the data.

History of Bayes' Theorem

Thomas Bayes (1701-1761), English Reverend[3]

[3] He never published what would eventually become his most famous accomplishment; his notes were edited and published after his death by Richard Price.

Bayes' Theorem for Classification

- The Bayes' theorem states that

$$\begin{aligned}\Pr(Y=k|X=x) &= \frac{\Pr(X=x,Y=k)}{\Pr(X=x)} = \frac{\Pr(Y=k)\Pr(X=x|Y=k)}{\Pr(X=x)} \\ &= \frac{\Pr(Y=k)\Pr(X=x|Y=k)}{\sum_{l=1}^{K}\Pr(Y=l)\Pr(X=x|Y=l)} =: \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}.\end{aligned}$$

- $f_k(x) = \Pr(X=x|Y=k)$ is the density function for X in class k. Here, we will use normal densities for these, separately in each class.[4]
- $\pi_k = \Pr(Y=k)$ is the marginal or prior probability for class k.
- $p_k(x) := \Pr(Y=k|X=x)$ is the posterior probability that an observation $X=x$ belongs to the kth class.
- Instead of estimating $p_k(x)$ directly as in logistic regression, discriminant analysis estimates π_k and $f_k(x)$ separately and plugs them in the Bayes' formula for $p_k(x)$.
- Estimating π_k is usually trivial [see below], but estimating $f_k(x)$ is not, so is the focus of the following discussion.

[4] Even if some X_j is discrete, normality assumption works well.

Illustration: $p = 1$ and $K = 2$

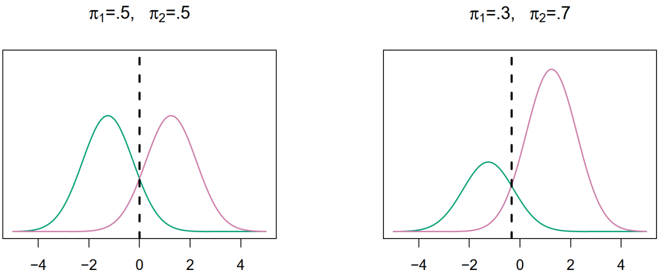


Figure: $\pi_k f_k(x)$, $k = 1, 2$

- Recall that the Bayes classifier (which has the lowest possible error rate out of all classifiers) classifies an observation x to the class for which $p_k(x)$ is largest.
- When $\pi_1 = \pi_2$, we classify a new point according to which density $f_k(x)$ is highest as shown in the left panel.
- When $\pi_1 \neq \pi_2$, we take them into account as well, and compare $\pi_k f_k(x)$. On the right, we favor the pink class – the decision boundary has shifted to the left.
- We will discuss three different estimates of $f_k(x)$ to approximate the Bayes classifer: LDA, QDA, and naive Bayes.

LDA for $p=1$

- To estimate $f_k(x)$, we make a simplifying assumption,

$$f_k(x)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)=:\phi\left(x|\mu_k,\sigma_k^2\right),$$

i.e., $X|(Y=k)\sim N\left(\mu_k,\sigma_k^2\right)$, where $\pi=3.14159\cdots$ is Archimedes' constant.

- LDA assumes further that $\sigma_1^2=\cdots=\sigma_K^2=\sigma^2$.
- Plugging this into Bayes' formula, we get

$$p_k(x)=\frac{\pi_k\phi\left(x|\mu_k,\sigma^2\right)}{\sum_{l=1}^K\pi_l\phi\left(x|\mu_l,\sigma^2\right)}.$$

- $\arg\max_k p_k(x)=\arg\max_k\log(p_k(x))$, so taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score:

$$\delta_k(x)=x\cdot\frac{\mu_k}{\sigma^2}-\frac{\mu_k^2}{2\sigma^2}+\log(\pi_k),\ \text{[Exercise]}$$

where as a function of x, $\delta_k(x)$ is called the discriminant function, which is linear in x so the name Linear DA.

Illustration: $p = 1$ and $K = 2$

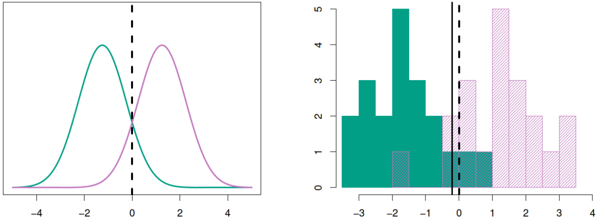


FIGURE 4.4. Left: *Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary.* Right: *20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.*

- Left: $\mu_1 = -1.25$, $\mu_2 = 1.25$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.
  - If $K = 2$ and $\pi_1 = \pi_2 = 0.5$, then the Bayes decision boundary is at $\{x|\delta_1(x) = \delta_2(x)\} = \left\{\frac{\mu_1+\mu_2}{2}\right\}$.
- Typically we don't know these parameters; we just have the training data. In that case LDA simply estimates the parameters and plug them into the rule.

Estimating the Parameters

- π_k can be estimated by
$$\hat{\pi}_k = \frac{n_k}{n},$$
where $n_k = \sum_{i=1}^{n} I(y_i = k)$ is the sample size of the kth class.
- μ_k and σ^2 can be estimated by
$$\begin{aligned}\hat{\mu}_k &= \frac{1}{n_k}\sum_{i:y_i=k} x_i,\\ \hat{\sigma}^2 &= \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2 = \sum_{k=1}^{K}\frac{n_k-1}{n-K}\cdot\hat{\sigma}_k^2\end{aligned}$$
is a weighted average of $\left\{\hat{\sigma}_k^2\right\}_{k=1}^{K}$ since $\sum_{k=1}^{K}\frac{n_k-1}{n-K}=1$, where $\hat{\sigma}_k^2 = \frac{\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2}{n_k-1}$ is the usual formula for the estimated variance in the kth class.
- Totally, we lose K degrees of freedom.
- The decision of LDA is based on $\arg\max_k \hat{\delta}_k(x)$, where $\hat{\delta}_k(x) = x\cdot\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$. [in the right panel of Figure 4.4, the decision boundary is $\frac{\hat{\mu}_1+\hat{\mu}_2}{2}$ since $n_1 = n_2$ such that $\hat{\pi}_1 = \hat{\pi}_2$]

LDA for $p > 1$

- Assume

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right),$$

i.e., $X|(Y=k) \sim N(\mu_k, \Sigma)$.

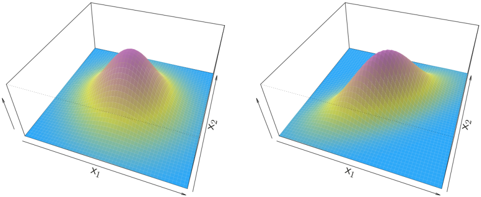


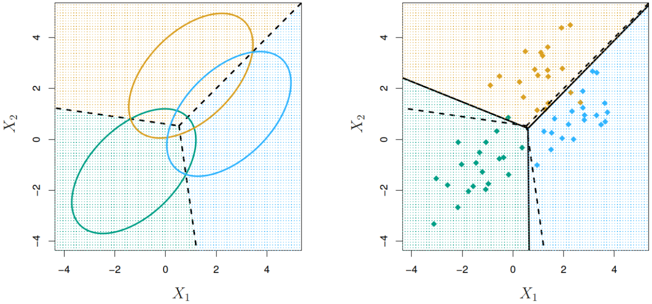
FIGURE 4.5. *Two multivariate Gaussian density functions are shown, with* $p = 2$. Left: *The two predictors are uncorrelated.* Right: *The two variables have a correlation of* 0.7.

- The discriminant function

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

is the vector/matrix version of the $p = 1$ case and is linear in x.

Illustration: $p = 2$ and $K = 3$



- Left: Ellipses that contain 95% of the probability for each of the three classes are shown. The dashed lines are the Bayes decision boundaries, where $\pi_1 = \pi_2 = \pi_3 = 1/3$. [Why linear? see the next slide]
- Right: 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines, where π_k, μ_k, and Σ are estimated similarly as in the $p = 1$ case.
- When $K > 3$, Appendix A shows how to visualize the discriminant rule efficiently.

More on the Decision Boundary

- The decison boundaries are determined by $\delta_k(x) = \delta_j(x)$ for some k and j.
- Because $\delta_k(x)$ is linear, $\{x | \delta_k(x) = \delta_j(x)\}$ is a point when $p = 1$, a straight line when $p = 2$, and a hyperplane when $p > 2$.
- In logistic regression, the decision boundaries are determined by

$$\{x | p_k(x) = p_j(x)\} = \{x | \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p\},$$

which is also linear in the X space, where

$$p_k(x) = \frac{e^{\beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0} + \beta_{l1}x_1 + \cdots + \beta_{lp}x_p}}$$

and recall that $\beta_K = \mathbf{0}$.

From $\delta_k(x)$ to Probabilities

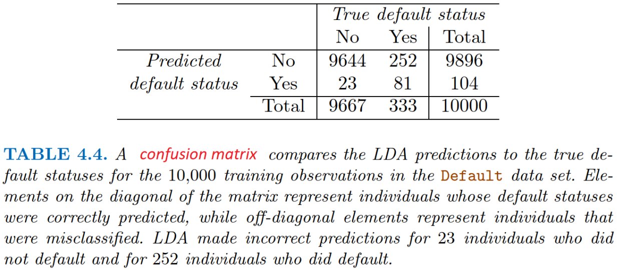
- Once we have estimates $\hat{\delta}_k(x)$, we can turn these into estimates for class probabilities:

$$\hat{p}_k(x) := \widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^K e^{\hat{\delta}_l(x)}}.$$

  - $\hat{p}_k(x) = e^{\hat{\delta}_k(x)+C} = e^{\hat{\delta}_k(x)}e^C$ for some constant C, so we need only determine e^C; while $\sum_{l=1}^K e^{\hat{\delta}_l(x)}e^C = 1$ so $e^C = \left(\sum_{l=1}^K e^{\hat{\delta}_l(x)}\right)^{-1}$.

- So classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\hat{p}_k(x)$ is largest.
- When $K = 2$, we classify to class 2 if $\hat{p}_2(x) \geq \hat{p}_1(x)$ or $\hat{p}_2(x) \geq 0.5$, else to class 1.

[Example] Credit Card Default



		True default status		
		No	Yes	Total
Predicted *default status*	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

TABLE 4.4. *A* confusion matrix *compares the LDA predictions to the true default statuses for the 10,000 training observations in the* `Default` *data set. Elements on the diagonal of the matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were misclassified. LDA made incorrect predictions for 23 individuals who did not default and for 252 individuals who did default.*

- The training error rate $= (23 + 252)/10000 = 2.75\%$, quite low.

Some Caveats:

- This is training error, and we may be overfitting. The higher p/n, the worse (see Section 6.4 in Lecture 4). Not a big concern here since $n = 10000$ and $p = 2$ [balance and student from logistic regression]!
- If we classified to the prior — always to class No in this case — we would make $333/10000$ errors, or only 3.33%.
- Of the true No's, we make $23/9667 = 0.2\%$ errors; of the true Yes's, we make $252/333 = 75.7\%$ errors!

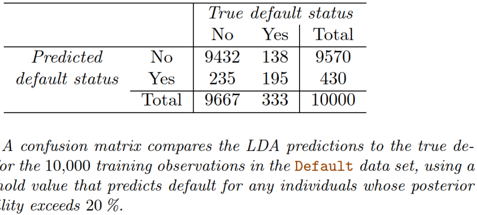
Types of Errors

- False positive rate: The fraction of negative examples that are classified as positive — 0.2% in example.
- False negative rate: The fraction of positive examples that are classified as negative — 75.7% in example.
- LDA tries to approximate the Bayes classifier which has the lowest total error rate, but does not distinguish the sources of errors.
- We produced Table 4.4 by classifying to class Yes if

$$\widehat{\Pr}\left(\text{default} = \text{Yes}|\text{balance}, \text{student}\right) \geq 0.5.$$

- We can change the two error rates by changing the threshold from 0.5 to some other value in $[0,1]$:

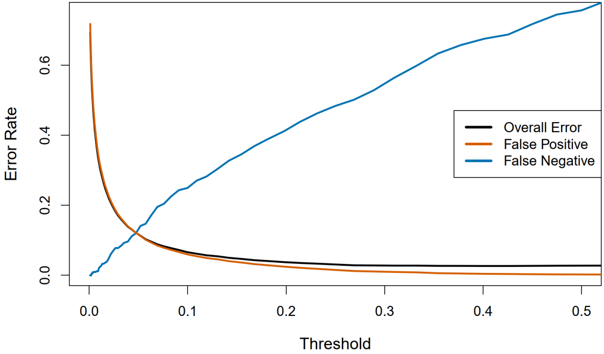
$$\widehat{\Pr}\left(\text{default} = \text{Yes}|\text{balance}, \text{student}\right) \geq \textit{threshold}.$$



		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

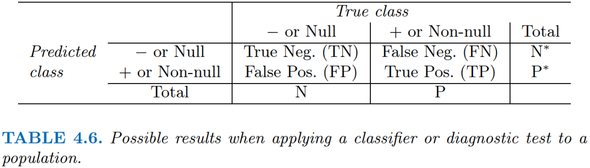
TABLE 4.5. *A confusion matrix compares the LDA predictions to the true default statuses for the* 10,000 *training observations in the* `Default` *data set, using a modified threshold value that predicts default for any individuals whose posterior default probability exceeds* 20 %.

[Example] Credit Card Default



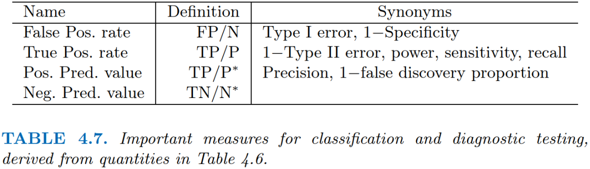
- The credit card company cares more about the false negative rate, the rate of identifying default as no default.
- In order to reduce this rate, we may want to reduce the threshold to 0.1 or less.
- Which threshold should be used? Use domain knowledge, such as the costs associated with the two types of errors.

Summary of Terms



		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N^*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P^*
	Total	N	P	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*



Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

- In Table 4.7, "false positive rate" is also called the false alarm rate, and the "true positive rate is also called the hit rate.
- In Table 4.7, note that the denominators for the first two terms are the actual population counts in each class, while the denominators for the last two terms are the total predicted counts for each class.
  - F_1 score$= 2\cdot\frac{prec\cdot rec}{prec+rec} = \frac{2}{\frac{1}{prec}+\frac{1}{rec}}$: overall summary that balances precision and recall.

ROC Curve

- The ROC curve displays both types of errors simultaneously for all possible thresholds.
  - "ROC", coming from communications theory, means receiver operating characteristics.
  - More properties of the ROC curve are explored in Assignment I.
- The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.

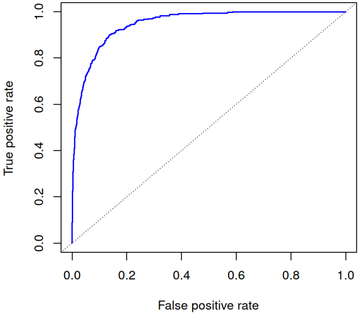


FIGURE 4.8. *A ROC curve for the LDA classifier on the* `Default` *data.*

AUC

- Sometimes we use the AUC (area under the curve) to summarize the overall performance for all possible thresholds.
- Higher AUC is good since it is the probability that a classifier will rank a randomly chosen $y = 1$ example higher than a randomly chosen $y = 0$ [see Appendix B].
- In Figure 4.8, AUC$= 0.95$, close to the maximum 1.
- We expect a classifier cannot perform worse than the diagonal line in Figure 4.8 whose AUC $= 0.5$.
- ROC curves are useful for comparing different classifiers.
  - The ROC curve for the logistic regression is almost the same as Figure 4.8, so logistic regression has the same AUC as LDA in this application.

Other Forms of Discriminant Analysis

- In the Bayes' formula of $p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$, we can get different classifiers by altering the forms for $f_k(x)$.
- With Gaussians but different Σ_k in each class, we get QDA, where the discriminant function
$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$$
is quadratic in x so the name Quadratic DA.
- Compared with LDA, QDA has a lower bias but higher variance; when p and/or K is large[5] and/or n is small, LDA is preferred since variance is the main concern now. [see Figure 4.9]
- With $f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)$ (conditional independence model – X_j's are independent given $Y = k$) in each class we get naive Bayes.
- For Gaussian this means the Σ_k's are diagonal.
- Compared with QDA, naive Bayes does not require normality assumption but assumes conditional independence.
- Many other forms, by proposing specific density models for $f_k(x)$, including non-parametric approaches.

[5] QDA has $Kp(p+1)/2 + Kp$ parameters while LDA has only $p(p+1)/2 + Kp$ parameters.

LDA versus QDA

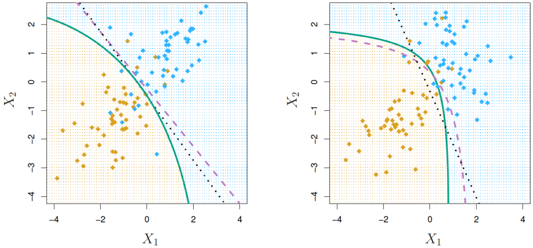


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$*. The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are as given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$*. Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

- Left: $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$.
- Right: $\begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} = \Sigma_1 \neq \Sigma_2 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$.

Naive Bayes

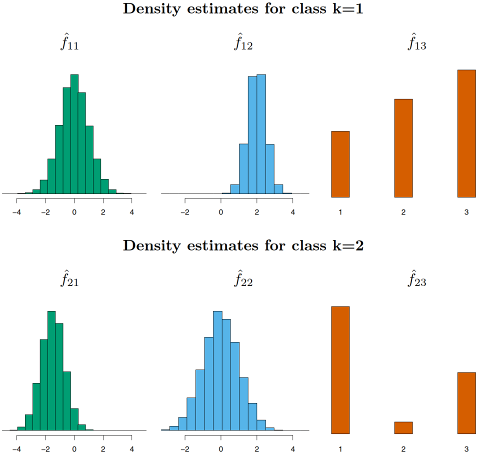
$$p_k(x) = \frac{\pi_k \times f_{k1}(x_1) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times \cdots \times f_{lp}(x_p)}$$

- Conditional independence is useful when p is large where multivariate methods like QDA and even LDA break down.
- Gaussian naive Bayes assumes each Σ_k is diagonal:

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\left[\frac{\left(x_j - \mu_{kj}\right)^2}{\sigma_{kj}^2} + \log\sigma_{kj}^2\right] + \log\pi_k .$$

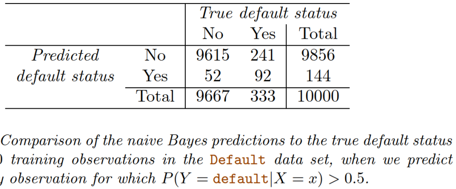
- Generally, for quantitative X_j, we can estimate $f_{kj}(x_j)$ by a histogram or a kernel density estimator which is a smoothed version of histogram.
- Naive Bayes can use for mixed feature vectors (qualitative and quantitative). If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function (histogram) over discrete categories.
- Despite strong assumptions (neglecting the association between X_j's), naive Bayes often produces good classification results, by decreasing variance although introducing some bias, especially when n is not large enough relative to p such that the joint distribution $f_k(x)$ cannot be effectively estimated.

Illustration: $p = 3$ and $K = 2$



- X_1 and X_2 are quantitative, and X_3 is qualitative with three levels.
- $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$. For $x^* = (0.4, 1.5, 1)^T$, $\hat{p}_1(x^*) = 0.944$ and $\hat{p}_2(x^*) = 0.056$.

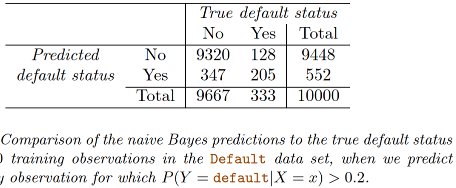
[Example] Credit Card Default



		True default status		
		No	Yes	Total
Predicted	No	9615	241	9856
default status	Yes	52	92	144
	Total	9667	333	10000

TABLE 4.8. *Comparison of the naive Bayes predictions to the true default status for the* 10, 000 *training observations in the* `Default` *data set, when we predict default for any observation for which* $P(Y = \texttt{default}|X = x) > 0.5$.

- Compared with Table 4.4, LDA has a slightly lower overall error rate, while naive Bayes (with Gaussian quantitative) correctly predicts a higher fraction of the true defaulters.



		True default status		
		No	Yes	Total
Predicted	No	9320	128	9448
default status	Yes	347	205	552
	Total	9667	333	10000

TABLE 4.9. *Comparison of the naive Bayes predictions to the true default status for the* 10, 000 *training observations in the* `Default` *data set, when we predict default for any observation for which* $P(Y = \texttt{default}|X = x) > 0.2$.

- Compared with Table 4.5, Naive Bayes has a higher error rate, but correctly predicts almost two-thirds of the true defaults. [$p = 2$ is not large relative to $n = 10,000$.]

A Comparison of Classification Methods

(Section 4.5)

Logistic Regression versus LDA/QDA

- $\arg\max_k p_k(x) = \arg\max_k \log\left(\frac{p_k(x)}{p_K(x)}\right)$, $k = 1, 2, \cdots, K$.
- One can show [read Section 4.5.1 for the details] that for LDA

$$\log\left(\frac{p_k(x)}{p_K(x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j,$$

  so it has the same form as logistic regression.
- The difference is in how the parameters are estimated.
  - Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as discriminative learning).
  - LDA uses the full likelihood based on $\Pr(X, Y)$ (known as generative learning).
  - Despite these differences, in practice the results are often very similar.
- For QDA,

$$\log\left(\frac{p_k(x)}{p_K(x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl}x_jx_l.$$

- Logistic regression can also fit quadratic boundaries like QDA, by explicitly including quadratic terms in the model.

Naive Bayes versus LDA/QDA

- For naitve Bayes,

$$\begin{aligned}\log\left(\frac{p_k(x)}{p_K(x)}\right) &= \log\left(\frac{\pi_k \prod_{j=1}^p f_{kj}(x_j)}{\pi_K \prod_{j=1}^p f_{Kj}(x_j)}\right)\\ &= \log\left(\frac{\pi_k}{\pi_K}\right)+\sum_{j=1}^p \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right) = a_k + \sum_{j=1}^p g_{kj}(x_j),\end{aligned}$$

  which takes the form of a generalized additive model [see Lecture 7].
- Any classifier with a linear decision boundary is a special case of naive Bayes with $g_{kj}(x_j) = b_{kj}x_j$.
  - LDA is a special case of naive Bayes although the former takes into account of the correlation among X_j while the latter does not.
- If $f_{kj}(x_j) = \phi\left(x_j|\mu_{kj},\sigma_j^2\right)$, where σ_j^2 does not depend on k, then Naive Bayes is a special case of LDA with $\Sigma =$diag$\left\{\sigma_1^2,\cdots,\sigma_p^2\right\}$.
- Neither QDA nor naive Bayes is a special case of the other – $\log\left(\frac{p_k(x)}{p_K(x)}\right)$ of the former can include interactions among x_j's while that of the latter can include more complex functions of x_j's than quadratic.

Summary

- Logistic regression is very popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Naive Bayes is useful when p is very large.
- None of these methods uniformly dominates the others: the choice of mothod depends on the true $f_k(x)$, and the relative magnitude of n and p which ties into the bias-variance trade-off.

(*) KNN versus the Methods in This Lecture

- Because KNN is completely nonparametric, it would dominate LDA and logistic regression when the decision boundary is highly non-linear, provided n is large and p is small.
- KNN requires a large n relative to p due to its nonparametric nature.
- When the decision boundary is non-linear, but n is moderate, or p is not very small, QDA may be preferred to KNN.
- Unlike logistic regression, KNN does not tell us which predictors are important.
- KNN is more nonparametric than naive Bayes, so suffers more from the curse of dimensionality.

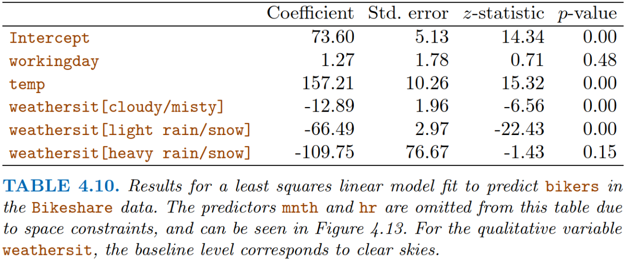
- Read also Section 4.5.2 for an empirical comparison of all these methods.

(**) Generalized Linear Models

(Section 4.6)

Linear Regression on the Bikeshare Data

- Sometimes, Y is neither quantitative nor qualitative, e.g., in the Bikeshare dataset, Y, the number of hourly users of a bike sharing program in Washington, DC. bikers, takes on non-negative integer values, or counts.



	Coefficient	Std. error	z-statistic	p-value
`Intercept`	73.60	5.13	14.34	0.00
`workingday`	1.27	1.78	0.71	0.48
`temp`	157.21	10.26	15.32	0.00
`weathersit[cloudy/misty]`	-12.89	1.96	-6.56	0.00
`weathersit[light rain/snow]`	-66.49	2.97	-22.43	0.00
`weathersit[heavy rain/snow]`	-109.75	76.67	-1.43	0.15

TABLE 4.10. *Results for a least squares linear model fit to predict* `bikers` *in the* `Bikeshare` *data. The predictors* `mnth` *and* `hr` *are omitted from this table due to space constraints, and can be seen in Figure 4.13. For the qualitative variable* `weathersit`*, the baseline level corresponds to clear skies.*

- In Table 4.10 and Figure 4.13, all coefficients of linear regression seem reasonable.

Problems with Linear Regression

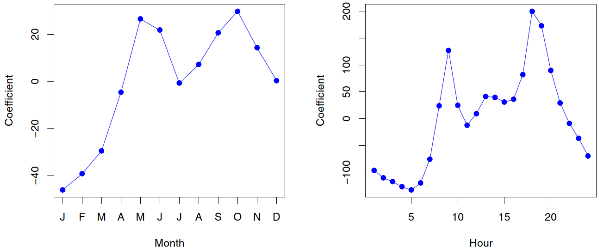


FIGURE 4.13. *A least squares linear regression model was fit to predict* `bikers` *in the* `Bikeshare` *data set.* Left: *The coefficients associated with the month of the year. Bike usage is highest in the spring and fall, and lowest in the winter.* Right: *The coefficients associated with the hour of the day. Bike usage is highest during peak commute times, and lowest overnight.*

- 9.6% of the fitted values, $\widehat{y}_i$, are negative.
- When the mean of bikers is small, its variance tends also to be small [see Figure 4.14], i.e., there is heteroskedasticity, but the linear regression model in Lecture 1 assumes homoskedasticity.
- In linear regression, Y should be continuous since the error term ε is continuous, but bikers is discrete.

Logarithmic Transformation on Y?

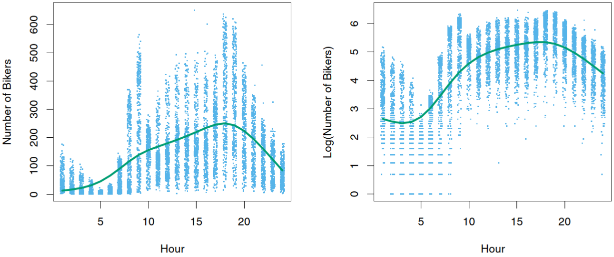


FIGURE 4.14. Left: *On the* Bikeshare *dataset, the number of bikers is displayed on the y-axis, and the hour of the day is displayed on the x-axis. Jitter was applied for ease of visualization. For the most part, as the mean number of bikers increases, so does the variance in the number of bikers. A smoothing spline fit is shown in green.* Right: *The log of the number of bikers is now displayed on the y-axis.*

- Fitting $\log(Y)$ on X can partially solve the first two problems. But it is hard to interpret $\hat{\beta}_j$; also, $\log 0$ is not well defined.

Poisson Regression on the Bikeshare Data

- Recall that, if $Y \sim$Poisson(λ), then

$$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, \cdots .$$

- $E[Y] = Var[Y] = \lambda$; a larger mean implies a larger variance.
- Given X, we model λ as a function of X, $\lambda(X)$, and

$$\log(\lambda(X)) = \beta_0 + \beta^T X,$$

or equivalently,

$$\lambda(X) = \exp\left(\beta_0 + \beta^T X\right) > 0.$$

- Estimate β_0 and β by maximum likelihood, where the likelihood function

$$\ell(\beta_0, \beta) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}.$$

- The results are qualitatively similar to those from linear regression. [see Table 4.11 and Figure 4.15]

	Coefficient	Std. error	z-statistic	p-value
Intercept	4.12	0.01	683.96	0.00
workingday	0.01	0.00	7.5	0.00
temp	0.79	0.01	68.43	0.00
weathersit[cloudy/misty]	-0.08	0.00	-34.53	0.00
weathersit[light rain/snow]	-0.58	0.00	-141.91	0.00
weathersit[heavy rain/snow]	-0.93	0.17	-5.55	0.00

TABLE 4.11. *Results for a Poisson regression model fit to predict* bikers *in the* Bikeshare *data.*

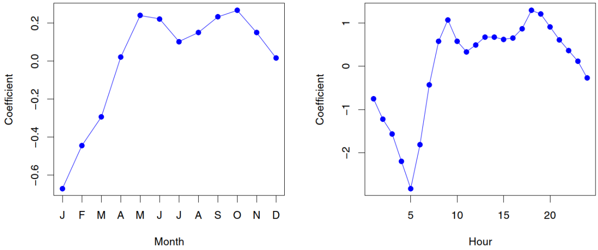


FIGURE 4.15. *A Poisson regression model was fit to predict* bikers *in the* Bikeshare *data set.*

Poisson Regression versus Linear Regression

- Interpretation: $\frac{E[Y|X_j=1]}{E[Y|X_j=0]} = \exp\left(\beta_j\right)$, so $\exp(-0.08) = 0.923$ means that on average, only 92.3% as many people will use bikes when it is cloudy relative to when it is clear.
- Mean-Variance Relationship: $E[Y|X] = \lambda(X) = Var[Y|X]$, so Poisson regression can handle (some specific form of) heteroskedasticity.
- Nonnegative Fitted Values: $\hat{y}_i = \hat{\lambda}(x_i) = \exp\left(\hat{\beta}_0 + \hat{\beta}^T X\right)$ cannot be negative.

Unifying Linear, Logistic and Poisson Regressions

- Linear regression, logistic regression, Poisson regression share some common properties:
- First, we use X to predict Y, and assume $Y|X$ follows some distributions – Gaussian, Bernoulli and Poisson.
- Second, we model $E\left[Y\right]$ as a function of X:

$$\begin{aligned}
\text{Linear Regression} &: \quad E\left[Y|X\right] = \beta_0 + \beta^T X, \\
\text{Logistic Regression} &: \quad E\left[Y|X\right] = \Pr\left(Y = 1|X\right) = \frac{e^{\beta_0+\beta^T X}}{1+e^{\beta_0+\beta^T X}}, \\
\text{Poisson Regression} &: \quad E\left[Y|X\right] = \lambda\left(X\right) = e^{\beta_0+\beta^T X}.
\end{aligned}$$

- All three $E\left[Y|X\right]$'s can be expressed using a link function, η,

$$\eta\left(E\left[Y|X\right]\right) = \beta_0 + \beta^T X,$$

where $\eta\left(\mu\right) = \mu$ for linear regression, $\eta\left(\mu\right) = \log\left(\mu/\left(1-\mu\right)\right) =$logit$\left(\mu\right)$ for logistic regression, and $\eta\left(\mu\right) = \log\left(\mu\right)$ for Poisson regression.

Generalized Linear Models

- These three distributions above are members of the exponential family.
- For other members of this family like exponential distribution, the Gamma distribution, and the negative binomial distribution, we can use different link functions to transform $E\left[Y|X\right]$ to a linear function of X.
- Any regression approach that follows this very general recipe is known as a generalized linear model (GLM).
- Other examples of GLM include Gamma regression and negative binomial regression, where

$$
\begin{aligned}
\eta\left(\mu\right) &= -\frac{1}{\mu} \text{ for Gamma regression,}^{6} \\
\eta\left(\mu\right) &= \log\left(\frac{\mu}{r+\mu}\right) \text{ for negative binomial regression.}
\end{aligned}
$$

- The negative binomial distribution has a pmf $\Pr\left(Y=k\right)=\binom{k+r-1}{r-1}p^{r}\left(1-p\right)^{k}$, $k=0,1,\cdots$, which is the probability for k failures until a predefined number r of successes have occurred in a sequence of independent Bernoulli trials; since $E\left[Y\right]=\left(1-p\right)r/p\leq\left(1-p\right)r/p^{2}=Var\left[Y\right]$ with equality as $p\to 1$, it can model the overdispersion phenomenon that the Poisson distribution cannot.

[6] The exponential distribution is a special case of the Gamma distribution, so shares the same link function.

Lab: Classification Methods (Section 4.7)

- The Stock Market Data
- Logistic Regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes
- K-Nearest Neighbors

Appendix A: Fisher's Discriminant Plot

(Section 4.3.3 of ESL)

Fisher's Discriminant Plot

- When there are K classes, linear discriminant analysis can be viewed exactly in a $K-1$ dimensional plot.
- Why? Because it essentially classifies to the closest centroid, and they span a $K-1$ dimensional plane.
  - This is convenient when $p \gg K$.
- Even when $K > 3$, we can find the "best" 2-dimensional plane for visualizing the discriminant rule.
- Fisher defined "best" to mean that the projected centroids were spread out as much as possible in terms of variance.
- So we try to find $Z_\ell = v_\ell^T X$, $\ell = 1, \cdots, \min(K-1, p)$, such that v_1 maximizes the Rayleigh quotient,

$$\max_a \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a},$$

or equivalently,

$$\max_a a^T \mathbf{B} a \text{ subject to } a^T \mathbf{W} a = 1,$$

and v_2 solves

$$\max_{a_2} \frac{a_2^T \mathbf{B} a_2}{a_2^T \mathbf{W} a_2} \text{ subject to } a_2^T \mathbf{W} a_1 = 0,$$

etc., where **B** is the between-class covariance and **W** is the within-class covariance of X.

Why Imposing the Constraint $a^T \mathbf{W} a = 1$?

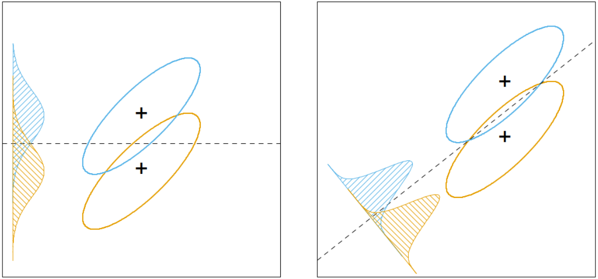
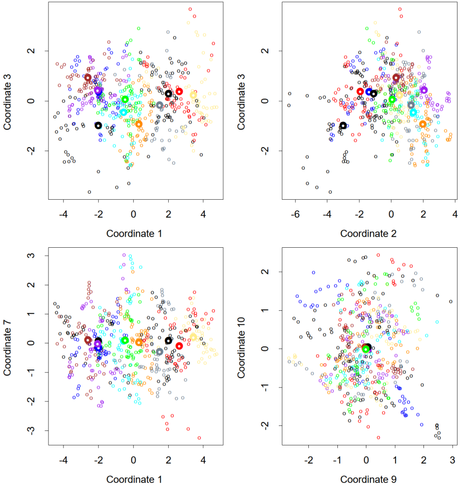


FIGURE *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

Illustration: $p = 10$ and $K = 11$



- As ℓ increases, the centroids become less spread out. In the lower right panel they appear to be superimposed, and the classes most confused.

Appendix B: More on Classification Assessment

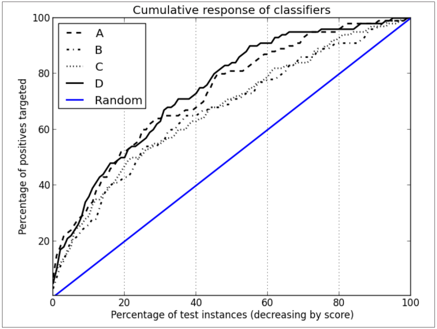
Interpretation of AUC

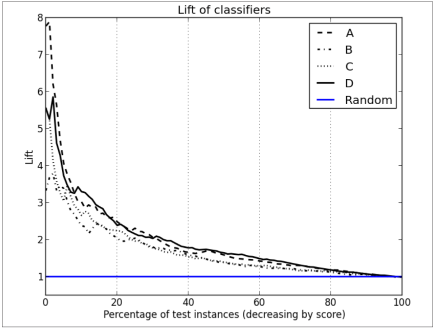
- For a generic score function $s(\cdot)$, we classify a test point x positive if $s(x)$ is greater than some threshold t.
  - E.g., in LDA, if we define class 2 as positive, then $s(\cdot) = \delta_2(\cdot) - \delta_1(\cdot)$.
- Denote $f_k(s)$, $k = 0, 1$, as the pdf of the scores for the negative and positive class, respectively, with cdf $F_k(s)$.
  - The pdf of all scores, $f(s)$, is $f_0(s)\pi_0 + f_1(s)\pi_1$.
- For the threshold t, the type I error is $1 - F_0(t)$, and the power is $1 - F_1(t)$.
- AUC is a plot of $1 - F_1(t)$ against $1 - F_0(t)$, so

$$\begin{aligned} \text{AUC} &= \int_0^1 (1 - F_1(t))\, d(1 - F_0(t)) \\ &= -\int_{\infty}^{-\infty} (1 - F_1(t)) f_0(t)\, dt \\ &= \int_{-\infty}^{\infty} (1 - F_1(t)) f_0(t)\, dt. \end{aligned}$$

which is the probability that a randomly drawn member of class 0 will produce a score lower than the score of a randomly drawn member of class 1.

Cumulative Response and Lift Curves





Continued

- y-axis in the cumulative response curve (CRC) is the percentage of positives that are correctly classified among all positives if all test instances below the threshold marked by the x-axis are classified as positive.
  - Why does the random classifier have the diagonal CRC?
  - What does the CRC for a perfect classifier look like?
- The lift curve is the ratio of the CRC and the diagonal line.
- Different from ROC where the tp/fp rate is computed using only the actual positive/negative examples, the x-axis of the CRC and lift curve involves both groups of examples, so they implicitly assume that the test set has exactly the same class priors as the population, but this need not be true especially when one class is rare in the population such that the other class is down-sampled (think about the responders of an advertisement).