Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 4: Classification (ISL ch. 4)

Instructor: Ma, Jun

Renmin University of China

October 28, 2021

Classification

- The linear regression model assumes that the response variable Y is quantitative: difference between two values is meaningful. Example: "Education" when measured in years.
- Qualitative variables take values (categories) in an unordered set $\mathscr{C}$: no natural ordering to the categories.
- Qualitative variables are also referred to as categorical variables.
  - `eye color` $\in \{$`brown`, `blue`, `green`$\}$
  - `email` $\in \{$`spam`, `ham`$\}$
- Predicting a qualitative response involves assigning the unseen response to a category, or class.
- Classification takes as input the feature vector X_0 and predicts its value Y_0: i.e., $C(X_0) \in \mathscr{C}$.

Classifiers

- Classification techniques are known as classifiers.
- Classifiers in this chapter estimate the probability $p_k(X_0) = \Pr[Y_0 = k \mid X_0]$ that the observation belongs to each of the categories $k \in \mathscr{C}$.
- Recall the Bayes optimal classifier:

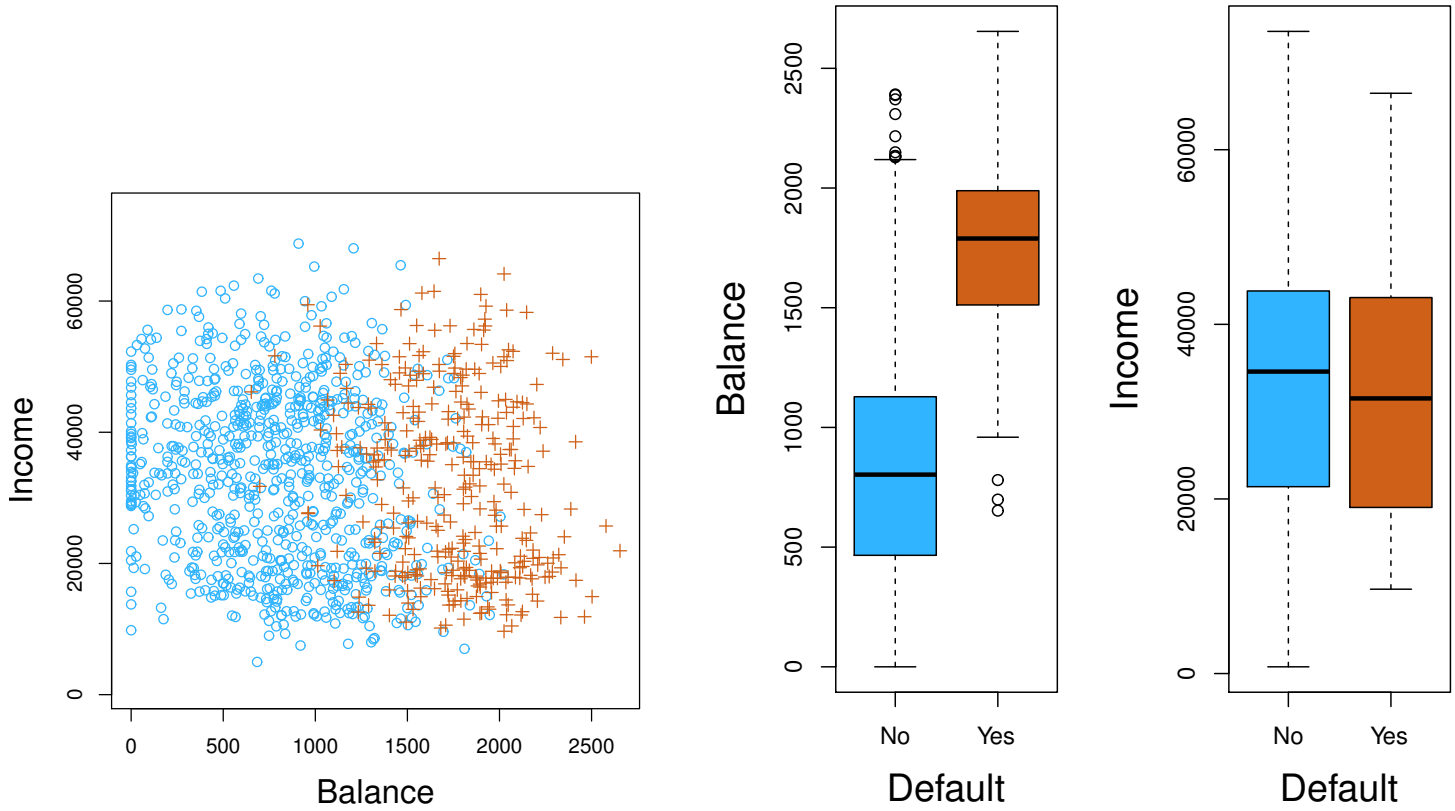
$$C(X_0) = \text{argmax}_j \, p_j(X_0)$$

  which satisfies:

$$\Pr[Y_0 \neq C(X_0) \mid X_0] \leq \Pr[Y_0 \neq g(X_0) \mid X_0]$$

  for any function g.
- We build a model for $p_k(x)$ and then approximate the Bayes optimal classifier.

The credit card default example



ISL Figure 4.1

- Predict whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance (the amount of money you owe the credit card company).
- The simulated `Default` data: those who defaulted in orange; those who did not in blue.

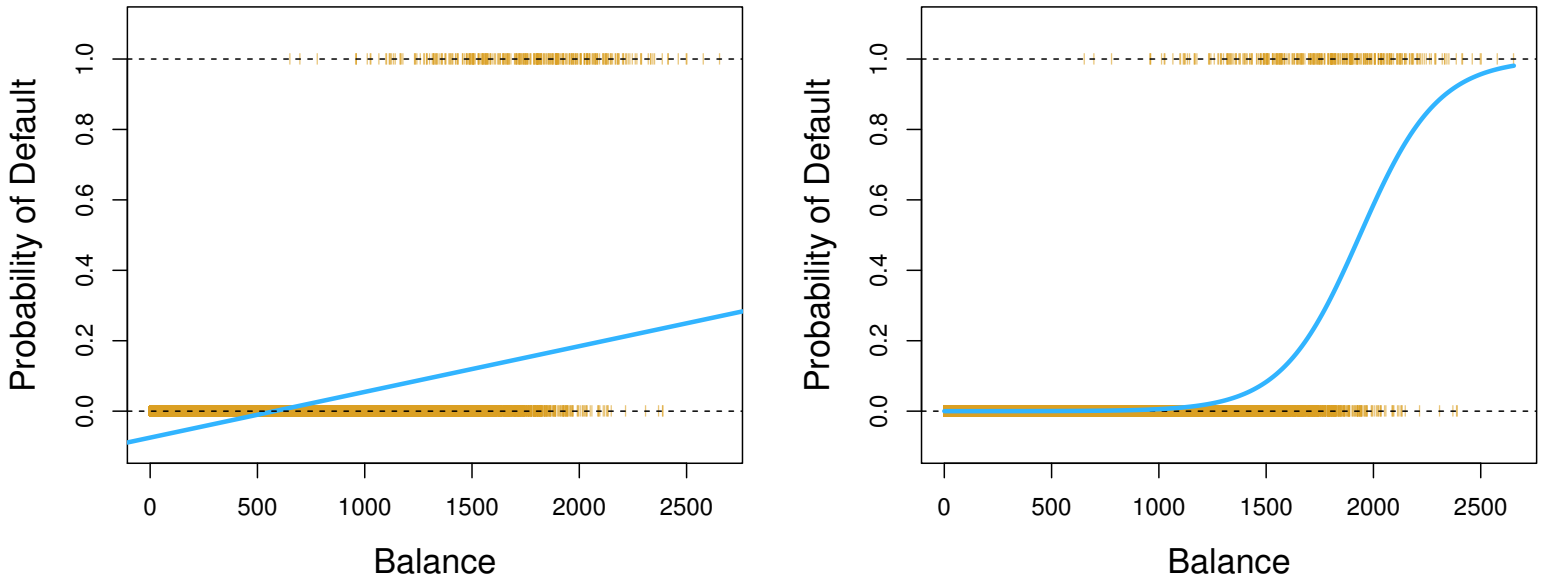
Why not linear regression?

- Linear regression mispecifies the conditional mean, if Y is binary: $\mathrm{E}[Y \mid X = x] = \Pr[Y = 1 \mid X = x]$ should be bounded and a predicted value from a linear regression can be bigger than 1 or smaller than 0.
- No natural way to convert a qualitative response if it takes more than two values which are unordered (predict the medical condition of a patient on the basis of her symptoms):

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure} \end{cases}$$

or

$$Y = \begin{cases} 1 & \text{if epileptic seizure;} \\ 2 & \text{if stroke;} \\ 3 & \text{if drug overdose.} \end{cases}$$



ISL 4.2

- 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 $\hat{Y} > 0.5$?
- It is clear that the in-sample mis-classification rate is high: all observations are classified as No.

Logistic regression

- Write $p(X) = \Pr[Y = 1 \mid X]$.
- Logistic regression specifies:

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

  for some parameter values (β_0, β_1).
- $p(X)$ has values between 0 and 1.

$$\frac{p(X)}{1 - p(X)} = \frac{\Pr[Y = 1 \mid X]}{\Pr[Y = 0 \mid X]}$$

  is called odds.
- The log odds is linear:

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

- In a linear regression model, β_1 gives the average change in Y associated with a one-unit increase in X.
- In a logistic regression model, increasing X by one unit changes the log odds by β_1.
- Take G to be the logitstic function:

$$G(z) = \frac{e^z}{1 + e^z}.$$

  This is the CDF for a standard logistic random variable.
- Partial effect of X on $p(X)$:

$$\frac{\mathrm{d}p(x)}{\mathrm{d}x} = G'(\beta_0 + \beta_1 x)\beta_1.$$

- The amount that $p(X)$ changes due to a one-unit change in X depends on the current value of X. But the sign of $p(X)$ is the same as that of β_1.

Maximum likelihood

- Let

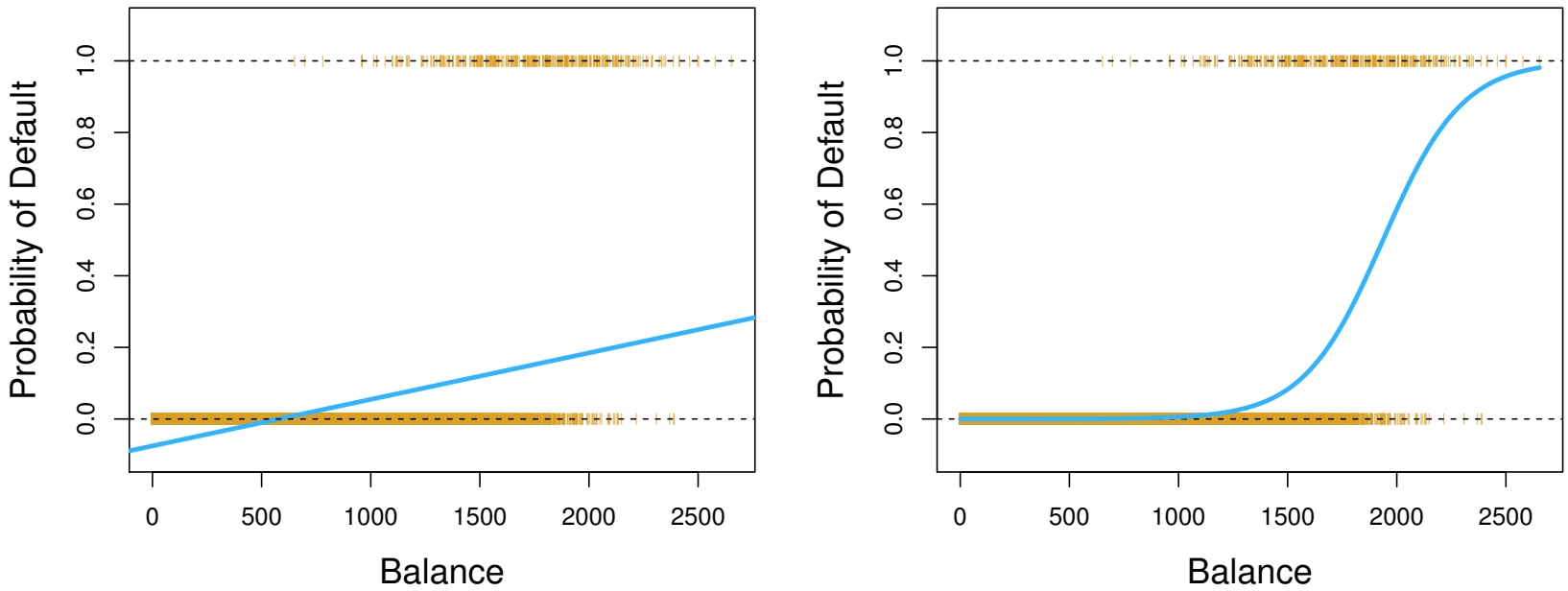
$$p\left(X_i; b_0, b_1\right) = \frac{e^{b_0 + b_1 X_i}}{1 + e^{b_0 + b_1 X_i}}.$$

- This likelihood gives the probability of the sample:

$$\ell\left(b_0, b_1\right) = \prod_{i=1}^{n} p\left(X_i; b_0, b_1\right)^{Y_i} \left(1 - p\left(X_i; b_0, b_1\right)\right)^{1 - Y_i}.$$

- We pick (b_0, b_1) to maximize the likelihood.

	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



ISL Figure 4.2

Predicted probabilities

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

$$\hat{p}\left(X\right) = \frac{e^{\hat{\beta}+\hat{\beta}X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006.$$

- With a balance of \$2000?

$$\hat{p}\left(X\right) = \frac{e^{\hat{\beta}+\hat{\beta}X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586.$$

A random utility model

- Econometricians' view of the logistic regression: it can be motivated by a random utility model.
- The unobserved utility of an agent Y^* is generated by

$$Y^* = \beta_0 + \beta_1 X + \epsilon,$$

  where ϵ is independent of X and has CDF G.
- We observe $Y = 1\left[Y^* > 0\right]$. The agent chooses $Y = 1$ if his or her net utility from doing so is positive and chooses $Y = 0$ otherwise.
- Then,

$$\begin{aligned}
\Pr\left[Y = 1 \mid X\right] =& \Pr\left[Y^* > 0 \mid X\right] \\
=& \Pr\left[\epsilon > -\left(\beta_0 + \beta_1 X\right) \mid X\right] \\
=& 1 - G\left(-\left(\beta_0 + \beta_1 X\right)\right) \\
=& G\left(\beta_0 + \beta_1 X\right).
\end{aligned}$$

- We observe the explanatory variable X and the actual choices Y from n independent agents.

Logistic regression with several variables

- Specify

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}$$

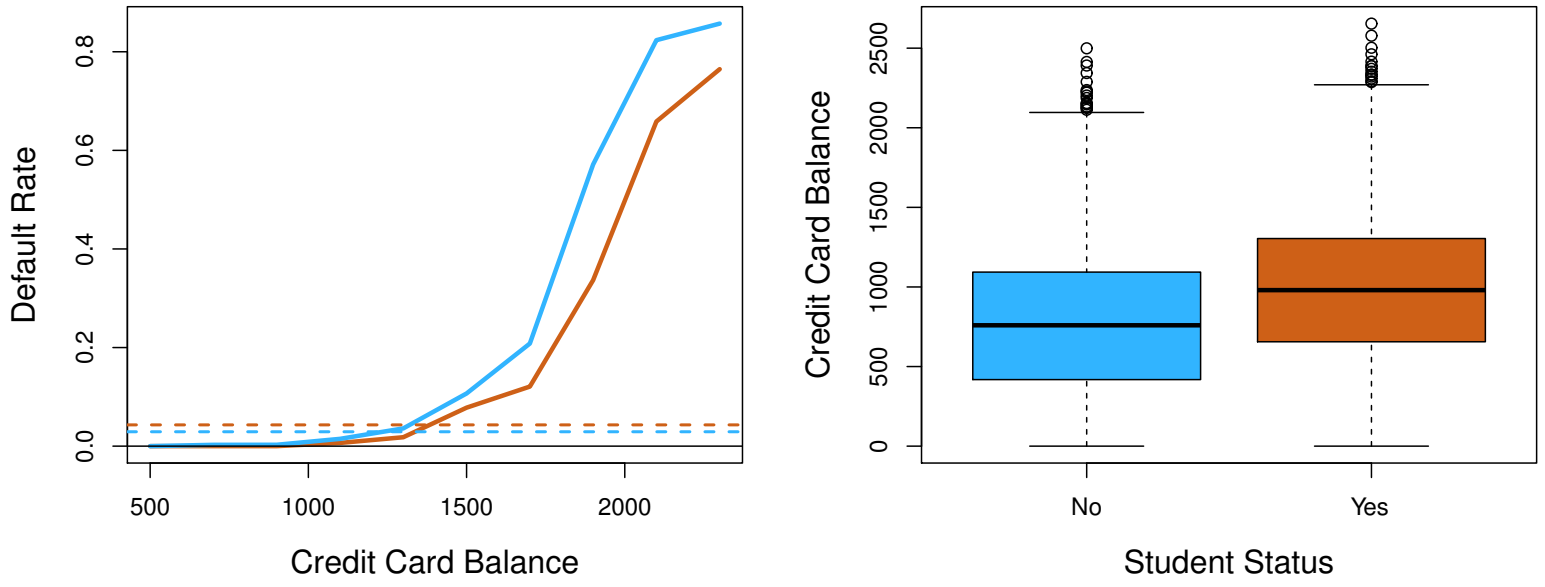
and then,

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X_1+\cdots\beta_p X_p.$$

	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

	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

- The coefficient for `student` in the single logistic regression is positive.
- The negative coefficient for `student` in the multiple logistic regression indicates that for a fixed value of balance and income, a student is less likely to default than a non-student.



ISL Figure 4.3

- Students tend to have higher balances than non-students.
- But for each level of `balance`, students default less than non-students.
- `Balance` has a positive effect on `Default`. Similar to omitted variable bias in linear models: single regression overestimates the effect of `student`.

Multinomial logistic regression

- $\mathscr{C} = \{0, 1, ..., K\}$ has more than two classes. $X \in \mathbb{R}^{p+1}$: p features and an intercept. $\beta_k \in \mathbb{R}^{p+1}$: coefficients ($k = 1, ..., K$).
- Specify:

$$\Pr[Y = k \mid X] = p_k(X) = \frac{e^{X^\top \beta_k}}{1 + \sum_{\ell=1}^{K} e^{X^\top \beta_\ell}}.$$

- Response probabilities should be summed up to 1:

$$\Pr[Y = 0 \mid X] = p_0(X) = \frac{1}{1 + \sum_{\ell=1}^{K} e^{X^\top \beta_\ell}}.$$

- Then, $\log(p_k(X)/p_0(X)) = e^{X^\top \beta_k}$: β_k are the marginal effects of X on the log-odds of k relative to the base category 0.
- Discriminant analysis is more suitable when $K > 1$.

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 $\Pr[Y \mid X]$.
- When we use normal distributions for each class, this leads to linear or quadratic discriminant analysis.

Bayes theorem

- Continuous (X, Y):

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int f_{X|Y}(x \mid y) f_Y(y) \, \mathrm{d}y},$$

  where $\int f_{X|Y}(x \mid y) f_Y(y) \, \mathrm{d}y = f_X(x)$.

- Discrete (X, Y):

$$\Pr[Y = k \mid X = x] = \frac{\Pr[X = x \mid Y = k] \cdot \Pr(Y = k)}{\sum_{k=1}^{K} \Pr[X = x \mid Y = k] \cdot \Pr(Y = k)}$$

  where $Y \in \{1, ..., K\}$ and

$$\sum_{k=1}^{K} \Pr[X = x \mid Y = k] \cdot \Pr[Y = k] = \Pr[X = x].$$

Linear discriminant analysis (LDA) for two classes

- Specify:

$$X \mid Y = 0 \sim \mathrm{N}\left(\mu_0, \Sigma\right)$$
$$X \mid Y = 1 \sim \mathrm{N}\left(\mu_1, \Sigma\right).$$

- By the (more general) Bayes theorem,

$$\Pr\left[Y = 1 \mid X\right] = \frac{\pi_1 f_1\left(X\right)}{\pi_0 f_0\left(X\right) + \pi_1 f_1\left(X\right)}$$
$$\Pr\left[Y = 0 \mid X\right] = \frac{\pi_0 f_0\left(X\right)}{\pi_0 f_0\left(X\right) + \pi_1 f_1\left(X\right)},$$

  where $\pi_k = \Pr\left[Y = k\right]$ and f_k is the conditional PDF of X given $Y = k$, $k \in \{0, 1\}$.

- The marginal distribution of Y (π_0, π_1) is left unspecified. (π_0, π_1) are easily estimated by sample averages.
- Estimation of (f_0, f_1) reduces to estimation of (μ_0, μ_1, Σ), which does not require numerical maximization (maximum likelihood).

More than two classes

- Similarly, $X \mid Y = k \in \{0, 1, ..., K\} \sim \mathrm{N}(\mu_k, \Sigma)$. Note that we assume the variances are the same.
- Note that in applications, X may have discrete variables like student status. The normality assumption is clearly violated but should be interpreted as a convenient model assumption.
- Then,

$$p_k(x) = \Pr[Y = k \mid X = x] = \frac{\pi_k f_k(x)}{\sum_{\ell=0}^{K} \pi_\ell f_\ell(x)},$$

  where $\pi_k = \Pr[Y = k]$ and f_k is the conditional PDF of X given $Y = k$, $k \in \{0, 1, ..., K\}$.
- We easily estimate f_k and π_k and get

$$\widehat{p}_k(x) = \frac{\widehat{\pi}_k \widehat{f}_k(x)}{\sum_{\ell=0}^{K} \widehat{\pi}_\ell \widehat{f}_\ell(x)}.$$

- LDA classifies an newly received observation X_0 to the class $\mathrm{argmax}_k \widehat{p}_k(X_0)$.

Why discriminant analysis?

- When the classes are well-separated, logistic regression model has bad performance. Discriminant analysis works better in these cases.
- If p is small and the distribution of the predictors X is approximately normal in each of the classes, discriminant analysis works better.
- Discriminant analysis is popular when we have more than two response classes.

LDA for $p = 1$

- The normal density has the form

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

  where μ_k is the mean and σ_k^2 is the variance (in class k).
- We assume that all the $\sigma_k^2 = \sigma^2$ are the same.
- Then,

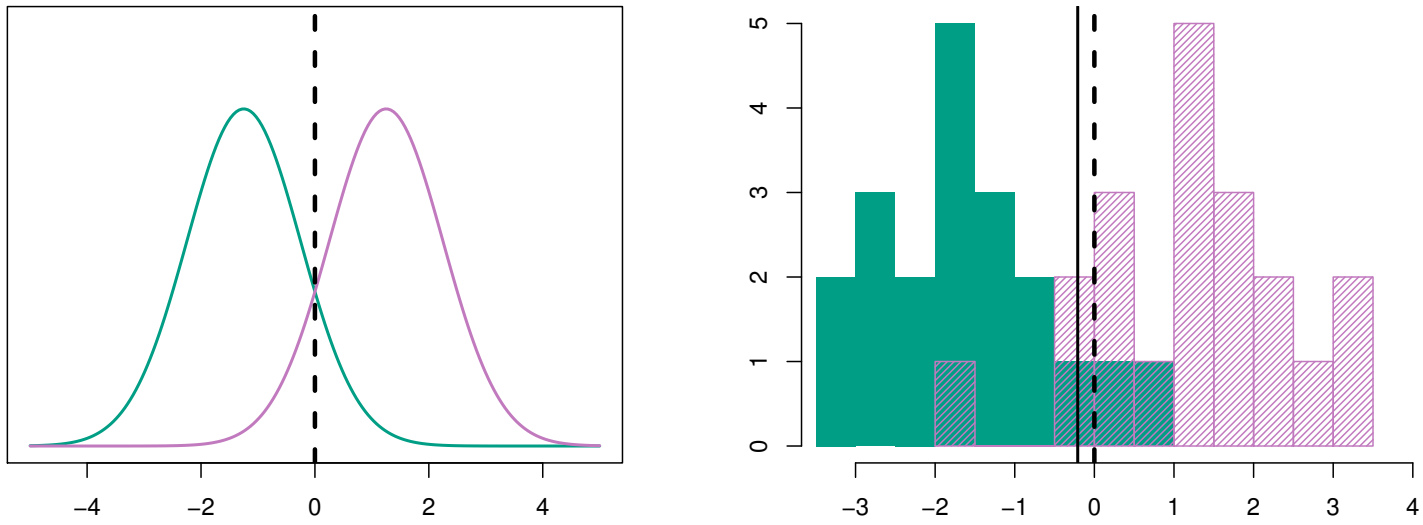
$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{\ell=0}^{K} \pi_\ell \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_\ell}{\sigma}\right)^2}}.$$

- To classify at the value $X_0 = x$, we need to see which of the $p_k(x)$ is largest.
- Taking logs, and discarding terms that do not depend on k, $C(x) = \text{argmax}_k p_k(x) = \text{argmax}_k \delta_k(x)$, where

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

  is known as the discriminant score.
- Note that by definition, $C(x)$ is the Bayes classifier, which yields the fewest misclassification errors, among all possible classifiers.
- Note that $\delta_k(x)$ is a linear function of x.
- If $\mathscr{C} = \{0, 1\}$ and $\pi_0 = \pi_1$, then $\delta_1(x) > \delta_0(x)$ if and only if $2x(\mu_1 - \mu_0) > \mu_1^2 - \mu_0^2$ and the decision boundary is at

$$x = \frac{\mu_1 + \mu_2}{2}.$$



ISL Figure 4.4

- Example with $\mu_1 = -1.5$ and $\mu_0 = 1.5$, $\pi_0 = \pi_1 = 0.5$ and $\sigma^2 = 1$.
- Dashed vertical line: the Bayes decision boundary.
- Solid vertical line: LDA decision boundary estimated from the training data.

Estimating the parameters

- Estimator of π_k:

$$\hat{\pi}_k = \frac{n_k}{n},$$

  where n_k is the number of training observations in the k-th class.

- Estimator of μ_k:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n} 1\left(Y_i = k\right) X_i,$$

  average of all the training observations from the k-th class.

- Estimator of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-K-1} \sum_{k=0}^{K} \sum_{i=1}^{n} 1\left(Y_i = k\right) \left(X_i - \hat{\mu}_k\right)^2 = \sum_{k=0}^{K} \frac{n_k - 1}{n-K-1} \cdot \hat{\sigma}_k^2$$

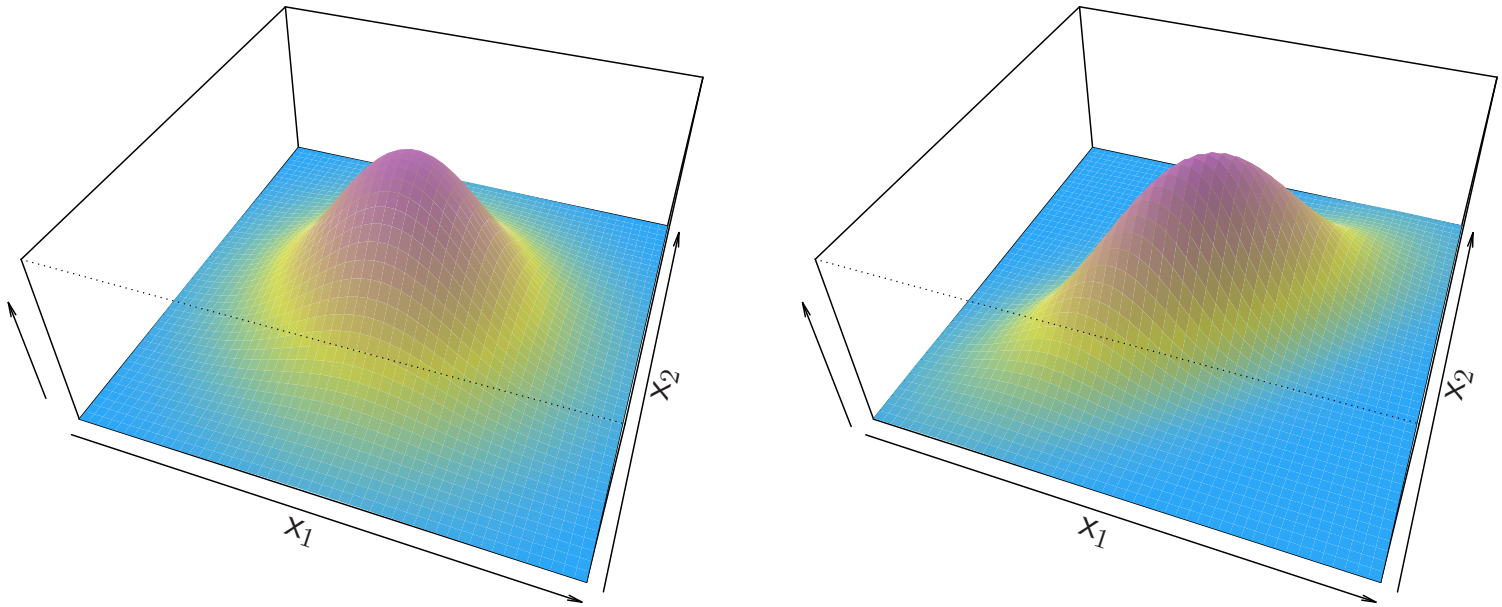
$$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n} 1\left(Y_i = k\right) \left(X_i - \hat{\mu}_k\right)^2 .$$

- $\hat{\sigma}^2$ is a weighted average of the sample variances for each of the $K+1$ classes.
- The LDA classifier assigns an observation $X_0 = x$ to the class for which

$$\hat{\delta}_k\left(x\right) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log\left(\hat{\pi}_k\right)$$

  is largest.

LDA for $p > 1$



ISL Figure 4.5

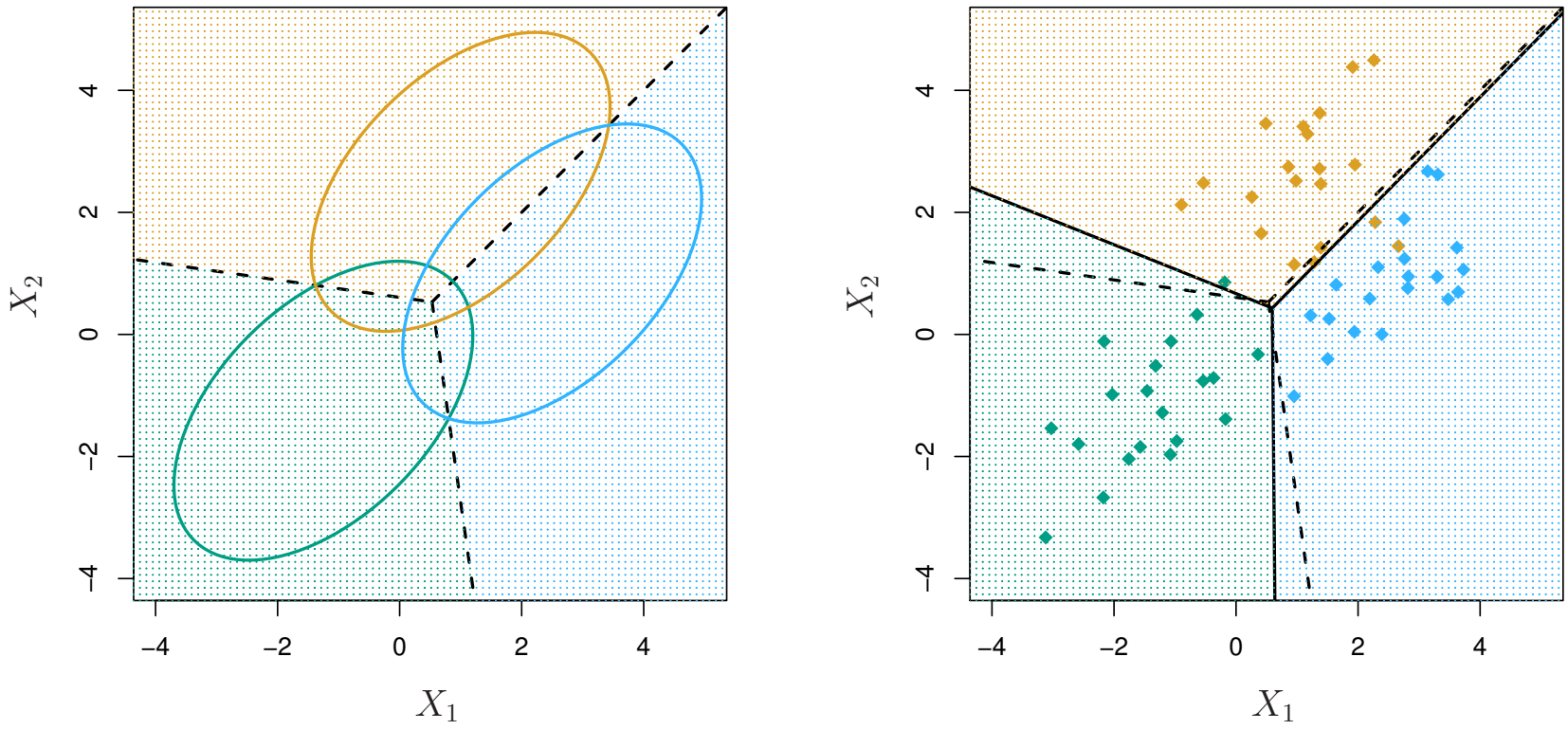
- Left: uncorrelated; right: correlation of 0.7.
- Multivariate normal density:

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

- In the case of $p > 1$ predictors, the LDA classifier assumes that the observations in the k-th class are drawn from a multivariate normal distribution $\mathrm{N}(\mu_k, \Sigma)$, where μ_k is a class-specific mean vector, and Σ is a covariance matrix that is common to all K classes.
- The Bayes classifier assigns an observation $X = x$ to the class for which

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

  is largest.
- We need to estimate the unknown parameters $\mu_0, \mu_1, ..., \mu_K$, $\pi_0, \pi_1, ..., \pi_K$ and Σ. LDA plugs these estimates to obtain quantities $\hat{\delta}_k(x)$, and classifies to the class for which $\hat{\delta}_k(x)$ is largest.
- Note that in $\delta_k(x)$ is a linear function of x. The LDA decision rule depends on x only through a linear combination of its elements.



ISL Figure 4.6

- $p = 2$, three classes with $\pi_0 = \pi_1 = \pi_2 = 1/3$.
- Dashed lines: the Bayes decision boundaries.

From $\delta_k(x)$ to probabilities

- Once we have estimates $\delta_k(x)$, we can turn these into estimates for class probabilities:

$$\hat{p}_k(x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{\ell=0}^{K} e^{\hat{\delta}_\ell(x)}}.$$

- Classifying to the largest $\delta_k(x)$ amounts to classifying class for which $p_k(x)$ is largest.

The credit card example

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Perform LDA on the `Default` data in order to predict whether or not an individual will default on the basis of credit card balance and student status.
- LDA results in a training error rate of $(23 + 252)/10000 = 2.75\%$.
- This is training error and we may be overfitting.
- If we classified to the prior (always to class `No`), we would make $333/10000 = 3.33\%$ error rate, only a bit higher than the LDA error rate.

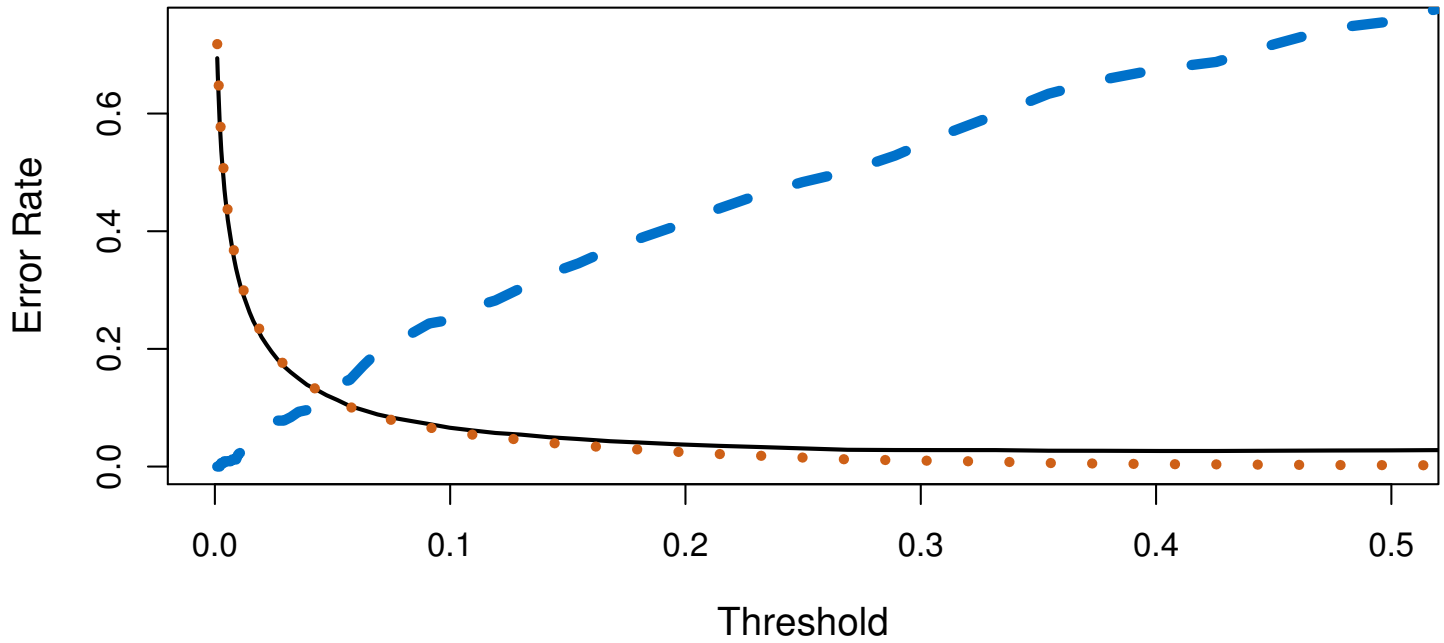
- Two types of errors:
  - incorrectly assign an individual who defaults to the no default category
  - incorrectly assign an individual who does not default to the default category.
- Only $23/9667 = 0.2\%$ of the individuals who did not default were incorrectly labeled.
- However, of the 333 individuals who defaulted, $252/333 = 75.7\%$ were incorrectly labeled by LDA.
- While the overall error rate is low, the error rate among individuals who defaulted is very high.

- False positive rate: the fraction of negative examples that are classified as positive, 0.2% in this example.
- False negative rate: the fraction of positive examples that are classified as negative, 75.7% in this example.
- In this example, the credit card company cares more about the false negative rate.
- LDA classifies to class `Yes` if

$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes} \mid \texttt{Balance}, \texttt{Student}) \geq 0.5.$$

- LDA is trying to approximate the Bayes classifier, which has the lowest total error rate out of all classifiers.
- We can change the two error rates by classifying with some Threshold $\in [0, 1]$:

$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes} \mid \texttt{Balance}, \texttt{Student}) \geq \text{Threshold}.$$



ISL Figure 4.7

- Black solid: the overall error rate; blue dashed: the fraction of defaulting customers that are incorrectly classified; orange dotted: the fraction of errors among the non-defaulting customers.
- In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less.

Logistic regression versus LDA

- Logistic regression:
  - Model the conditional distribution $Y \mid X$.
  - The distribution of X is not modeled.
  - Use MLE to estimate. This requires numerical optimization.
  - Economic justification: random utility model.
- LDA:
  - Model the conditional distribution $X \mid Y$.
  - The distribution of Y is not modeled.
  - Estimation: sample means, variances, and covariances of X.
  - No clear economic model.

Quadratic discriminant analysis (QDA)

- In the LDA model,

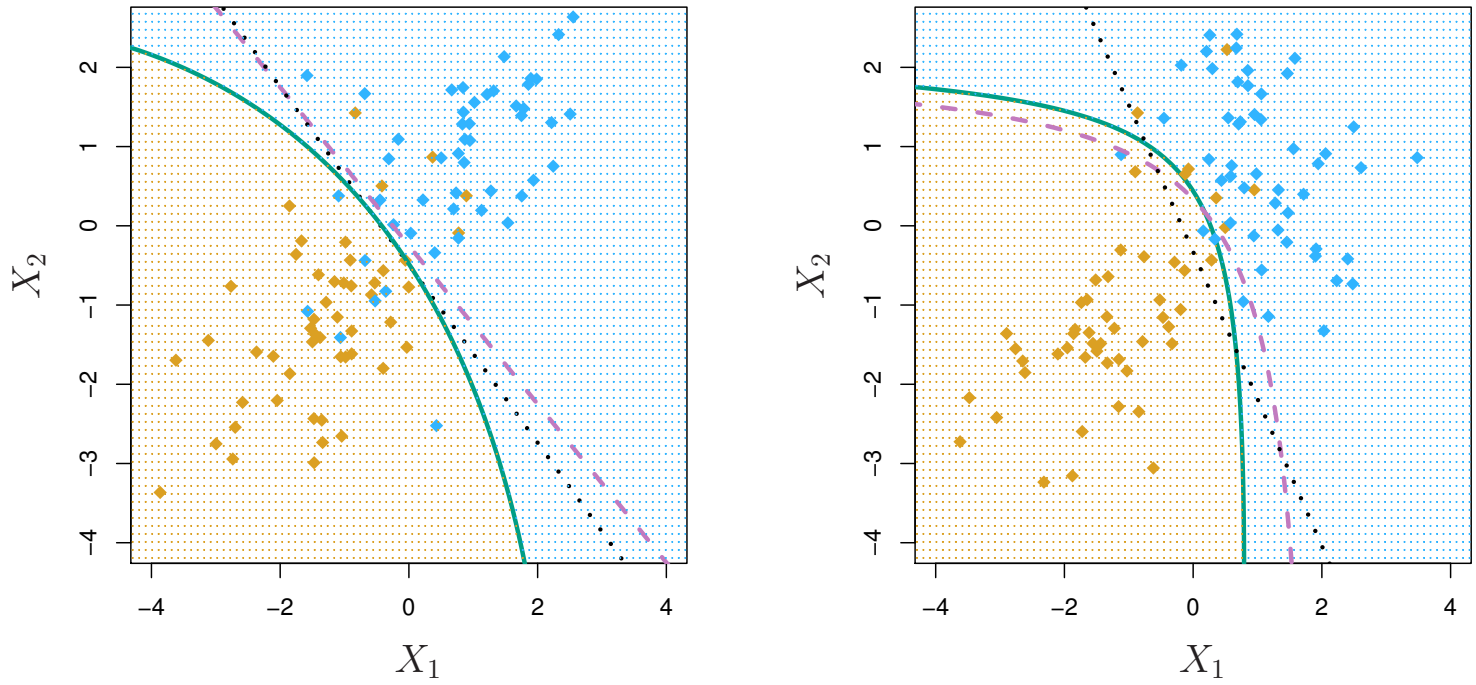
$$p_k(x) = \Pr[Y = k \mid X = x] = \frac{\pi_k f_k(x)}{\sum_{\ell=0}^{K} \pi_\ell f_\ell(x)},$$

  f_k are normal densities, with the same covariance matrix Σ in each class.
- QDA: different Σ_k in each class.
- The Bayes optimal classifier assigns $X_0 = x$ to the class for which

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

  is largest.$\delta_k(x)$ is now a quadratic function of x.

- When there are p predictors, then estimating a covariance matrix requires estimating $p(p+1)/2$ parameters. QDA estimates a separate covariance matrix for each class, for a total of $(K+1)\,p(p+1)/2$ parameters.
- LDA assumes that the $K+1$ classes share a common covariance: $(K+1)\,p$ linear coefficients to estimate.
- LDA is much less flexible than QDA, and so has substantially lower variance. LDA can suffer from high bias, if the $K+1$ conditional distributions do not have similar conditional variances.



ISL Figure 4.9

- Left: Bayes optimal (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with $\Sigma_0 = \Sigma_1$. The Bayes decision boundary is linear, so LDA works better.
- Right: $\Sigma_0 \neq \Sigma_1$. Since the Bayes decision boundary is non-linear, QDA works better.

Naive Bayes

- In

$$p_k(x) = \Pr[Y = k \mid X = x] = \frac{\pi_k f_k(x)}{\sum_{\ell=0}^{K} \pi_\ell f_\ell(x)},$$

  $f_k(x)$ is a p-dimensional conditional PDF for the k-th class.
- In general, estimating a p-dimensional density function is challenging. LDA and QDA take a parametric approach.
  - LDA assumes that f_k is the density function for a multivariate normal random variable with class-specific mean μ_k, and common covariance Σ.
  - QDA allows for class-specific covariance Σ_k.
  - When p is large relatively to n, QDA and LDA break down.
- The naive Bayes classifier assumes $f_k = \Pi_{j=1}^{p} f_{kj}$ (features are independent) in

$$p_k(x) = \Pr[Y = k \mid X = x] = \frac{\pi_k f_k(x)}{\sum_{\ell=0}^{K} \pi_\ell f_\ell(x)}.$$

- f_{kj}: the conditional PDF of X_j given $Y = k$.

- The independence assumption is interpreted as a model assumption for convenience, rather than what we actually believe in.
- The naive Bayes assumption introduces some bias, but reduces variance, leading to a classifier that works quite well in practice as a result of the bias-variance trade-off.
- Gaussian naive Bayes assumes $X_j \mid Y = k \sim \mathrm{N}\left(\mu_{kj}, \sigma_{kj}^2\right)$ and each Σ_k is diagonal ($X_1, ..., X_p$ are independent):

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

- If X_j is qualitative, replace f_{kj} with its probability mass function.
- Naive Bayes often performs well in practical applications.