COM 525000 – Statistical Learning

Lecture 4 – Classification

Y.-W. Peter Hong

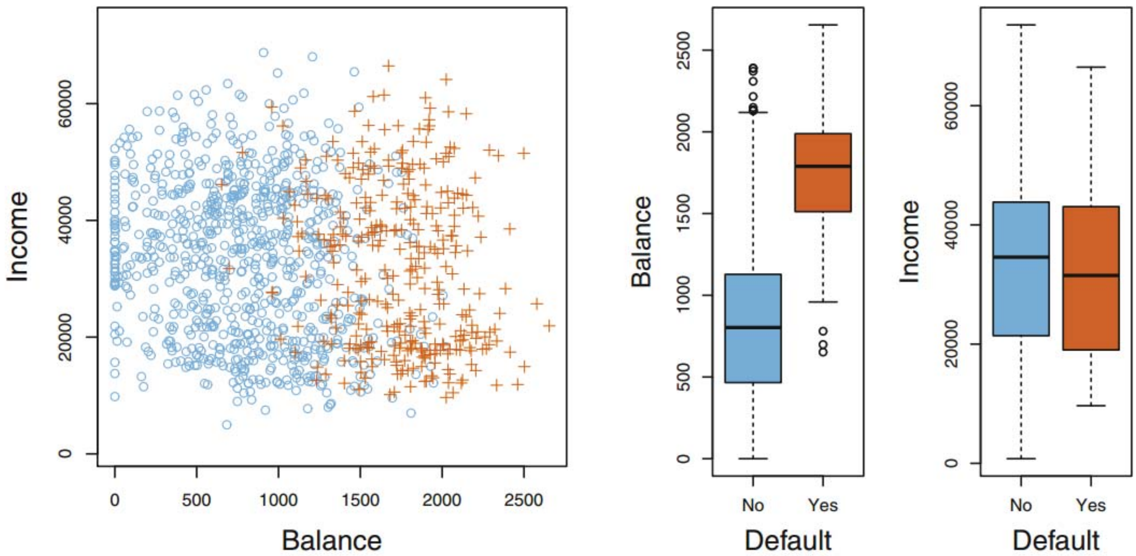
Classification

- **Classification** refers to the prediction of qualitative or categorical responses based on the observation.
  ➔ Often done by predicting the probability of each of the categories (similar to *regression*).
- Three widely-used classifiers:
  – **Logistic regression;**
  – **Linear discriminant analysis;**
  – **k nearest neighbors.**
- Other methods, such as generalized additive models (Chapter 7), trees, random forests, boosting (Chapter 8), and support vector machines (Chapter 9).

Motivating Example

Example: (Simulated Default Data Set)

- Annual income versus credit card balance of 10,000 individuals, and their default status.



Estimating Quantitative Values

- **Question:** Why not assign quantitative values to different categories and predict these values?

  E.g., for stroke, drug overdose, epileptic seizure, let

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

  and estimate the value of Y. (➔ implies ordering.)

- This is "ok" in the binary case, e.g.,

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

  ➔ Large and small predicted values represents two different directions (instead of an ordering).

Estimating Probabilities

- By using linear regression to predict Y, the predicted response can be given by

$$\hat{Y} = \begin{cases} 1, & \hat{f}(X) = X^T\beta \geq 0.5 \\ 0, & \hat{f}(X) = X^T\beta < 0.5. \end{cases}$$

- In estimation theory, the estimator that minimizes the squared loss function $L(Y, f(X)) = (Y - f(X))^2$ is

$$\begin{aligned} f^*(x) &= \arg\min_c E_{Y|X}\left[(Y-c)^2 | X = x\right] \\ &= E[Y|X = x] \\ &= \Pr(Y = 1 | X = x) \left(\triangleq p(x) \right) \end{aligned}$$

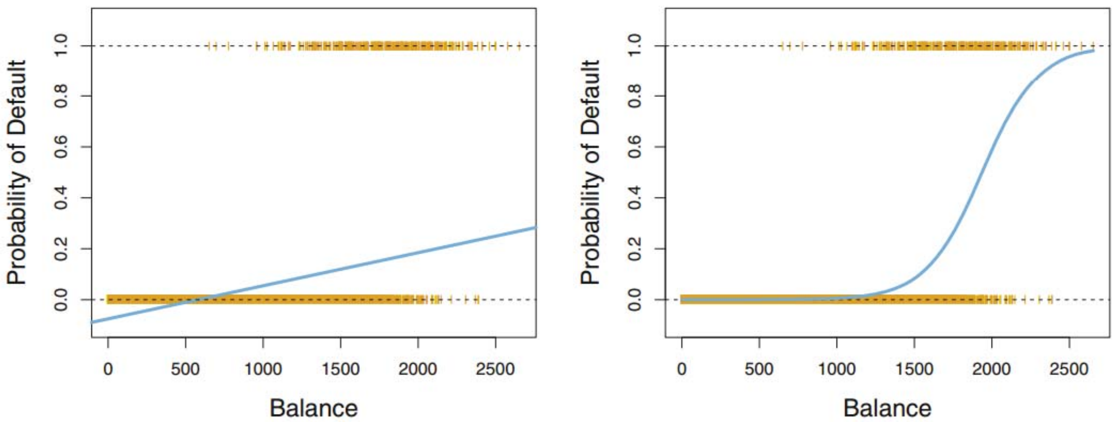
➔ The linear least squares solution $\hat{f}(X) = X^T\beta$ can be viewed as a linear approximation of this function.

Linear vs Logistic Model $(p = 1)$

- **Linear model:** $p(X) = \beta_0 + \beta_1 X$ (may go beyond [0,1])
- **Logistic model:**

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \quad \boxed{\Rightarrow p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}$$

➔Imposes a linear model on the *log-odds* (or *logit*).



Estimating β_0 and β_1 $(p = 1)$

- **Maximum likelihood method**: Choose β_0 and β_1 to maximize the likelihood that $y_1, \ldots, y_n$ are the responses given that the inputs are $x_1, \ldots, x_n$, i.e.,

$$\begin{aligned}\ell(\beta_0, \beta_1) &= \prod_{i=1}^{n} \Pr(Y = y_i | X = x_i; \beta_0, \beta_1) \\ &= \prod_{i:y_i=1} p(x_i; \beta_0, \beta_1) \prod_{i:y_i=0} (1 - p(x_i; \beta_0, \beta_1)).\end{aligned}$$

➔

Logistic Regression on the Default Data Set

	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

- E.g., the estimated probability of default for an individual with balance $1000 is

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

  ➔ For balance $2000, this estimated prob. is 0.586.

- For qualitative variable, such as “student”, we can define dummy variable

$$X = \begin{cases} 1, & \text{if student,} \\ 0, & \text{if not student.} \end{cases}$$

Multiple Logistic Regression (1/2)

- For the case with multiple predictors $X_1, X_2, \ldots, X_p$,

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

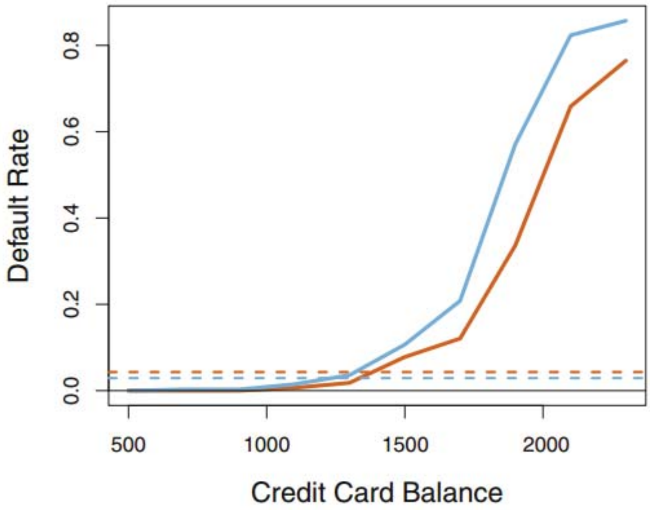
Example:

	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

➔ Student less likely to default for *fixed* balance and income, but more likely to default on the average.

Multiple Logistic Regression (2/2)



— students

— non-students

Estimating the Coefficients

- By adopting the *maximum likelihood method*, the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ are chosen to maximize

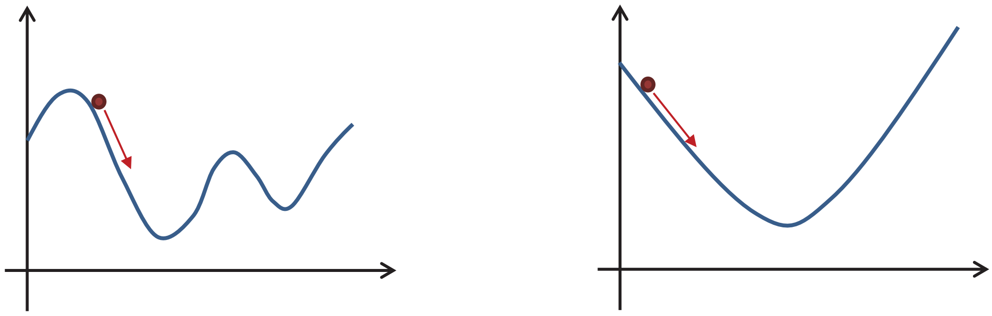
$$\log \ell(\beta) = \sum_{i=1}^{n} \left[y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta)) \right]$$
$$= \sum_{i=1}^{n} \left[y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right]$$
$$= -\sum_{i=1}^{n} \log\left(1 + e^{-(2y_i - 1)\beta^T x_i}\right)$$

➔ concave and, thus, has a unique global maximum.

- Numerical methods, e.g., gradient descent, iteratively reweighted least squares (IRLS).

Gradient Descent (1/3)

- Gradient descent is like having a ball roll down a hilly surface, e.g., $J(\beta) = \sum_{i=1}^{n} \log\left(1 + e^{-(2y_i - 1)\beta^T x_i}\right)$.



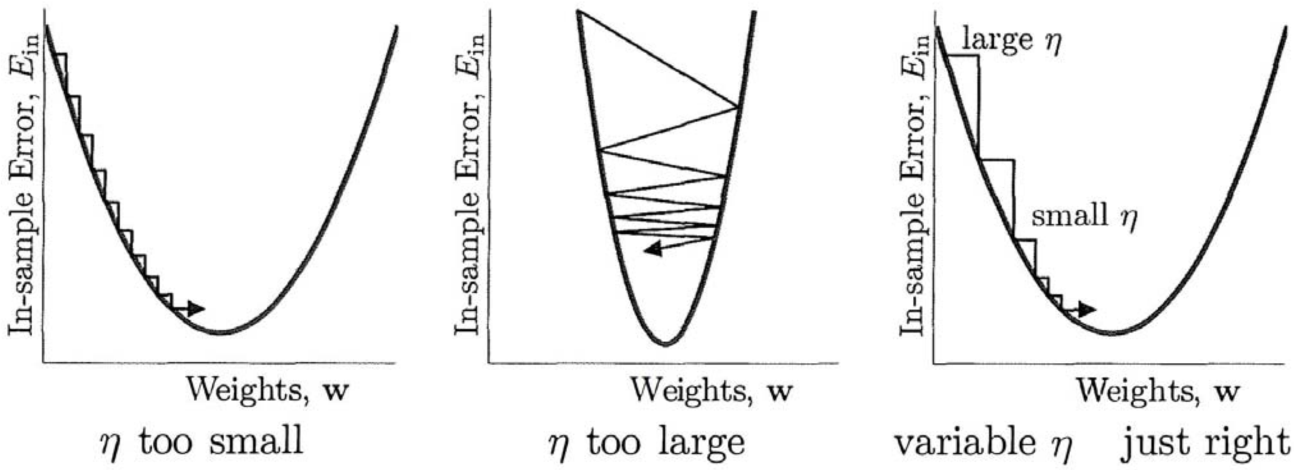
1. Start at random point $\beta(0)$.
2. In iteration $k+1$, take a step along the steepest slope

$$\beta(k+1) = \beta(k) - \eta \nabla J(\beta(k))$$

where $\nabla J(\beta(k)) = -\sum_{i=1}^{n} \frac{(2y_i - 1)x_i}{1+e^{-(2y_i-1)\beta(k)^T x_i}}$.

Gradient Descent (2/3)

Gradient Descent (3/3)



Iteratively Reweighted Least Squares (1/2)

- Newton-Raphson algorithm uses update method

$$\beta(k+1) = \beta(k) - \left(\frac{\partial^2 \log \ell(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial \log \ell(\beta)}{\partial\beta}\Big|_{\beta=\beta(k)}$$

- The first derivative is given by

$$\frac{\partial \log \ell(\beta)}{\partial\beta} = \sum_{i=1}^{n} x_i(y_i - p(x_i;\beta)) = \mathbf{X}^T(\mathbf{y} - \mathbf{p}(\beta))$$

where $\mathbf{p}(\beta) = [p(x_1;\beta), \ldots, p(x_n;\beta)]^T$ and the second derivative or Hessian matrix is

$$\frac{\partial^2 \log \ell(\beta)}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} x_i x_i^T p(x_i;\beta)(1 - p(x_i;\beta)) = -\mathbf{X}^T\mathbf{W}(\beta)\mathbf{X}$$

where $\mathbf{W}(\beta) = \mathrm{diag}(\mathbf{p}(\beta))(\mathbf{I} - \mathrm{diag}(\mathbf{p}(\beta)))$.

Iteratively Reweighted Least Squares (2/2)

- That is, we have

$$\begin{aligned}
&\beta(k+1)\\
&=\beta(k)+(\mathbf{X}^T\mathbf{W}(\beta(k))\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}(\beta(k)))\\
&=(\mathbf{X}^T\mathbf{W}(\beta(k))\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\beta(k))\left[\mathbf{X}\beta(k)+\mathbf{W}(\beta(k))^{-1}(\mathbf{y}-\mathbf{p}(\beta(k)))\right]\\
&=(\mathbf{X}^T\mathbf{W}(\beta(k))\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\beta(k))\mathbf{z}(k)
\end{aligned}$$

  where $\mathbf{z}(k) \triangleq \mathbf{X}\beta(k) + \mathbf{W}(\beta(k))^{-1}(\mathbf{y}-\mathbf{p}(\beta(k)))$ is the *adjusted response*.

- The update in each iteration is the solution to the weighted least squares problem:

$$\beta(k+1) \leftarrow \arg\min_{\beta}(\mathbf{z}(k)-\mathbf{X}\beta)^T\mathbf{W}(\beta(k))(\mathbf{z}(k)-\mathbf{X}\beta)$$

➔ **Iteratively reweighted least squares (IRLS)**

Multinomial Logistic Regression (1/2)

- Suppose that there are $K > 2$ classes (represented by the response variable $Y \in \{1,\ldots,K\}$).
- **Goal:** Model the log-posterior probabilities

$$\log p_k(x) = \log \Pr(Y=k|X=x),\ k=1,\ldots,K,$$

  as linear functions while satisfying $\sum_{k=1}^K p_k(x) = 1$.

- The model is given by

$$\log\left(\frac{p_1(x)}{p_K(x)}\right) = \beta_1^T x$$

$$\log\left(\frac{p_2(x)}{p_K(x)}\right) = \beta_2^T x$$

$$\vdots$$

$$\log\left(\frac{p_{K-1}(x)}{p_K(x)}\right) = \beta_{K-1}^T x$$

$$p_k(x) = \frac{e^{\beta_k^T x}}{1+\sum_{\ell=1}^{K-1} e^{\beta_\ell^T x}} \quad \text{for } k=1,\ldots,K-1$$

$$p_K(x) = \frac{1}{1+\sum_{\ell=1}^{K-1} e^{\beta_\ell^T x}}$$

Multinomial Logistic Regression (2/2)

- The coefficients are chosen to maximize

$$\log \ell(\beta) = \sum_{k=1}^{K} \sum_{i:y_i=k} \log p_k(x_i;\beta_k)$$

- Given coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_K$, where, $\hat{\beta}_K = 0$, the predicted output for new input x_0 is given by

$$\hat{Y}(x_0) = \arg \max_{k \in \{1,\ldots,K\}} \frac{e^{\hat{\beta}_k^T x_0}}{1 + \sum_{\ell=1}^{K-1} e^{\hat{\beta}_\ell^T x_0}}$$

$$\left(= \arg \max_{k \in \{1,\ldots,K\}} \Pr(Y = k | X = x_0) \right)$$

➔The (estimated) ***maximum a posterior probability (MAP)*** *detector*.

Bayes Theory for Classification (1/3)

- Suppose that there are K classes and, thus, the response variable $Y \in \{1, 2, \ldots, K\}$.
- Let $\pi_k = \Pr(Y = k)$ be the prior probability that $Y = k$, and let $f_k(x) \triangleq \Pr(X = x | Y = k)$ be the conditional density function of X given $Y = k$.
- A classifier can be described as

$$\hat{Y}(x) = \begin{cases} 1, & x \in \mathcal{R}_1 \\ 2, & x \in \mathcal{R}_2 \\ \vdots & \vdots \\ K, & x \in \mathcal{R}_K \end{cases}$$

where $\mathcal{R}_k$ is the decision region for class k. Here, $\mathcal{R}_k \cup \mathcal{R}_\ell = \emptyset$ and $\cup_{k=1}^{K} \mathcal{R}_k = \mathcal{X}$.

Bayes Theory for Classification (2/3)

- The probability of prediction error is
  $\Pr(\hat{Y}(X) \neq Y) =$

➔To minimize the error probability, we should let $x \in \mathcal{R}_k$ whenever $\Pr(Y = k|X = x) > \Pr(Y = \ell|X = x), \ \forall \ell \neq k.$

Bayes Theory for Classification (3/3)

- The prediction that minimizes the probability of classification error is given by

$$\begin{aligned}\hat{Y}(x) &= \arg \max_{k \in \{1,\ldots,K\}} \Pr(Y = k|X = x) \\ &= \arg \max_{k \in \{1,\ldots,K\}} \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \left(\triangleq p_k(x) \right).\end{aligned}$$

➔Logistic regression aims to estimate the log of the posterior probability using a linear model.

➔Why not estimate π_k and $f_k(x)$ directly?

Gaussian with Equal Variance Assumption

- *Estimating* π_k *is relatively easy, but estimating* $f_k(x)$ *can be more difficult unless assuming a simple form.*
- Suppose that $f_k(x)$ is assumed to be **Gaussian**, i.e.,

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

  with **equal variance** $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma^2$.
  (Recall that μ_k and σ_k^2 are the mean and variance parameters for the kth class.)
- In this case, we have

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

Bayes Classifier in the Gaussian Case (1/2)

- Under the Gaussian and equal variance assumption, the Bayes classifier yields
  $\hat{Y}(x) = \arg \max_{k \in \{1,\ldots,K\}} p_k(x) =$

➔ $\delta_k(x) \triangleq x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k$ is the classification criterion.

Bayes Classifier in the Gaussian Case (2/2)

- For $K = 2$ and $\pi_1 = \pi_2$, the Bayes classifier decides on class 1 if

  $\delta_1(x) > \delta_2(x) \ \Rightarrow$

  and decides on class 2, otherwise.
- The Bayes decision boundary is given by x such that $\delta_1(x) = \delta_2(x)$. That is,

Linear Discriminant Analysis for $p = 1$

- **Linear discriminant analysis (LDA)** assumes that $f_k(x)$ is Gaussian with equal variance and applies the Bayes classifier with estimates

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \ \hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2, \ \hat{\pi}_k = \frac{n_k}{n},$$

  where n_k is the number of observations in class k.
- This yields the approximated classification criterion

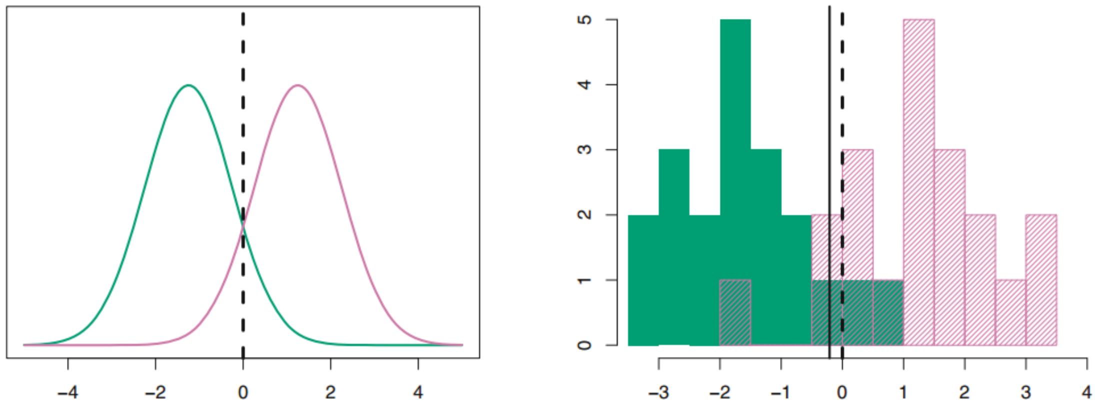
$$\hat{\delta}_k(x) = x \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

- In LDA, the classification is performed by taking

$$\hat{Y} = \arg \max_{k \in \{1,\dots,K\}} \hat{\delta}_k(x).$$

Example

- Left: the Gaussian density functions for two classes.
- Right: histogram of 20 random observations from each class (i.e., $n_1 = n_2 = 20$ and, thus, $\hat{\pi}_1 = \hat{\pi}_2$).



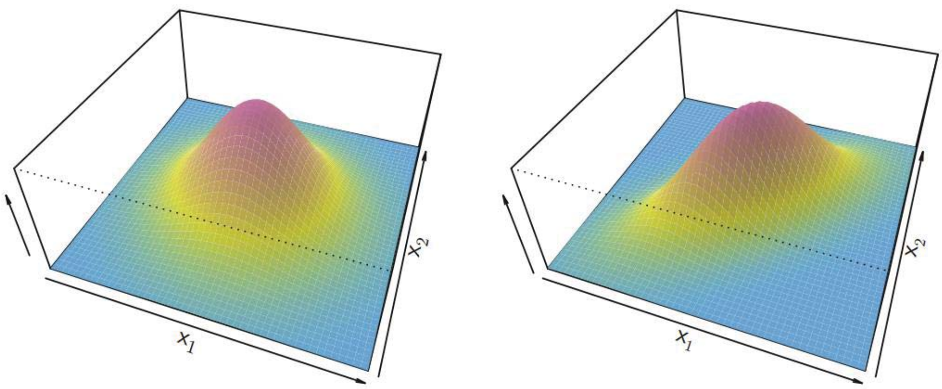
$$\hat{\delta}_1(x) \gtrless \hat{\delta}_2(x) \Rightarrow x\frac{\hat{\mu}_1}{\hat{\sigma}^2} - \frac{\hat{\mu}_1^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_1) \gtrless x\frac{\hat{\mu}_2}{\hat{\sigma}^2} - \frac{\hat{\mu}_2^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_2)$$

Multivariate Gaussian

- Consider the case with p predictors $X = (X_1, X_2, \ldots, X_p)^T$.
- The density function under class k is assumed to be multivariate Gaussian, i.e.,

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

where $\mu_k = E[X|Y=k]$, $\mathbf{\Sigma}_k = E[(x-\mu_k)(x-\mu_k)^T|Y=k]$.



Linear Discriminant Analysis for $p>1$

- For $\boldsymbol{\Sigma}_1=\cdots=\boldsymbol{\Sigma}_K=\boldsymbol{\Sigma}$, the Bayes classifier is

$$\hat{Y}(x)=\arg\max_{k\in\{1,\ldots,K\}} p_k(x)=\arg\max_{k\in\{1,\ldots,K\}}\delta_k(x)$$

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

- **LDA** approximates the Bayes classifier by replacing the parameters $\mu_k,\boldsymbol{\Sigma},\pi_k$ with their estimates

$$\hat{\mu}_k=\frac{1}{n_k}\sum_{i:y_i=k}x_i,\quad \hat{\boldsymbol{\Sigma}}=\frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T,$$

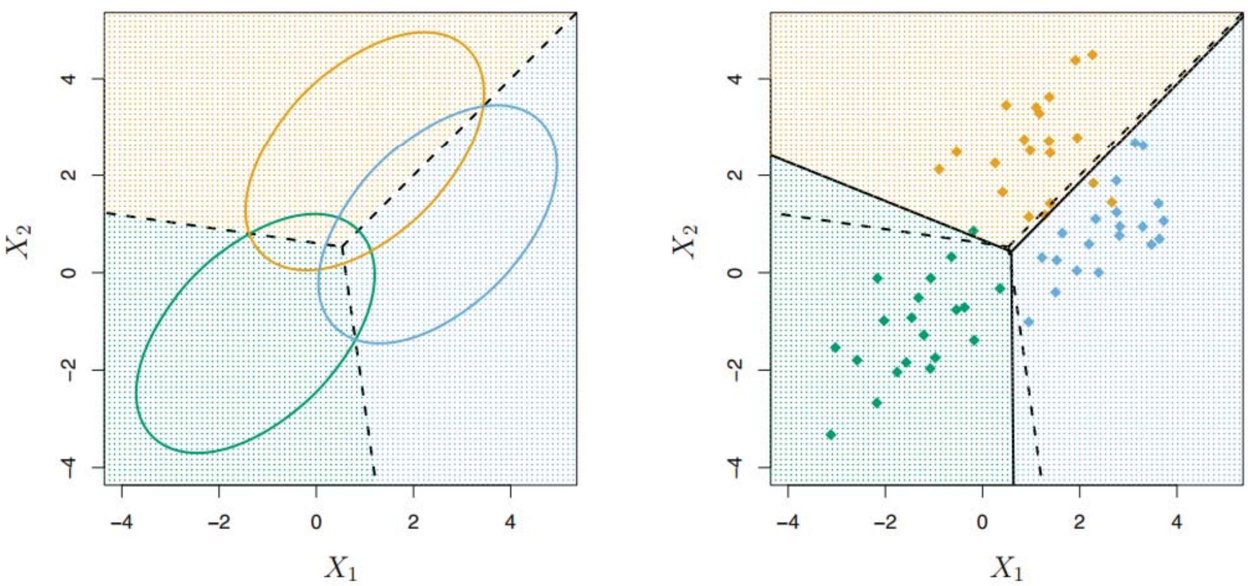
  and $\hat{\pi}_k=\frac{n_k}{n}$. That is, the classification is given by

$$\hat{Y}(x)=\arg\max_{k\in\{1,\ldots,K\}}\hat{\delta}_k(x)$$

$$=\arg\max_{k\in\{1,\ldots,K\}} x^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k-\frac{1}{2}\hat{\mu}_k^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mu}_k+\log\hat{\pi}_k.$$

Example

- The boundary between classes k and ℓ is given by the line $\delta_k(x)=\delta_\ell(x)$ for the Bayes classifier, and by the line $\hat{\delta}_k(x)=\hat{\delta}_\ell(x)$ for the LDA classifier.



Confusion Matrix (1/2)

- For the Default data set, fitting the LDA model to the 10,000 training data yields *training* error rate **2.75%**. (Here, the predictors are balance & student status.)

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

➔Training errors are usually less than test errors.

➔Only 3.33% of individuals in the data set default (i.e., a *null* classifier has only 3.33% error rate).

Confusion Matrix (2/2)

- *Sensitivity* (*specificity*) is the percentage of true (non-) defaulters that are correctly identified.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

- Only $\frac{81}{333} = 24.3\%$ of defaulters were correctly identified.
  ➔ sensitivity=24.3% (i.e., error rate in "default"=75.7%)
- $\frac{9644}{9667} = 99.8\%$ of non-defaulters were correctly identified.
  ➔ specificity=99.8% (i.e., error rate in "non-default"=0.2%)

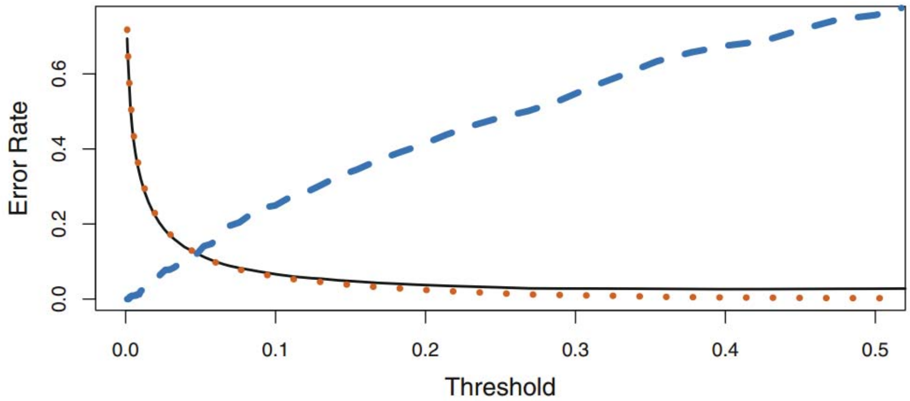
Remark: Bayes classifier minimizes the *average error rate*, not the individual error rates in different classes.

Average vs Conditional Error Rate

- Recall that, in the Default data set, we have

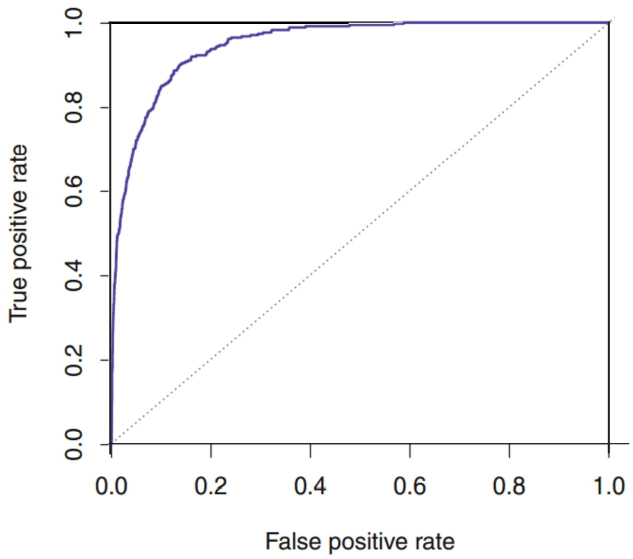
$$Y = \begin{cases} 1, & \text{if default} \\ 0, & \text{if no default.} \end{cases}$$

- The Bayes classifier yields $\hat{Y} = 1$ for input $X = x$ if
  $\Pr(Y=1|X=x) > \Pr(Y=0|X=x) = 1 - \Pr(Y=1|X=x)$
  $\Rightarrow \Pr(Y=1|X=x) > 0.5$



Receiver Operating Characteristics (ROC)

- The receiver operating characteristics (ROC) curve displays the true positive rate vs false positive rates.



➔The overall performance of a classifier can be measured by the *area under the (ROC) curve (AUC)*.

Quadratic Discriminant Analysis

- **Quadratic discriminant analysis (QDA)** is similar to LDA, but assuming a different variance for each class.
- That is, for class k, $X \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}_k)$.
- Under this assumption, the Bayes classifier is

$$\hat{Y}(x) = \arg \max_{k \in \{1,\ldots,K\}} \delta_k(x)$$

where

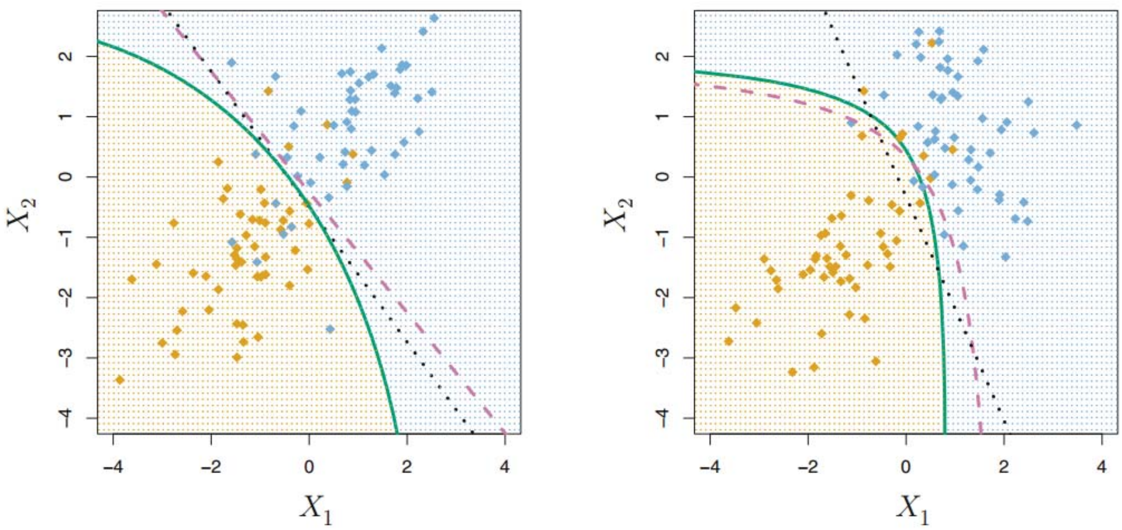
$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T \mathbf{\Sigma}_k^{-1}(x - \mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log \pi_k$$

➔ quadratic in terms of variable x.

- QDA is more flexible but has much more parameters to estimate. ➔ problematic with small training data

Example: LDA vs QDA

- Left: Two Gaussian classes have a common correlation of 0.7 among the predictors.
- Right: Orange class has correlation 0.7 and blue class has correlation -0.7.



Comparison of Classification Methods (1/3)

- Two-class case with $p_1(x)$ and $p_2(x)=1-p_1(x)$ being the prob. that $X=x$ belongs to classes 1 and 2.
- In logistic regression, we let

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \beta_0 + \beta_1 x$$

- In LDA, we assume that

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{2} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}, \ k=1,2$$

and, thus,

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\frac{\pi_1}{\pi_2} - \frac{\mu_1^2-\mu_2^2}{2\sigma^2} + \frac{\mu_1-\mu_2}{\sigma^2}x = c_0 + c_1 x.$$

➔ LR and LDA apply different fitting procedures.

Comparison of Classification Methods (2/3)

- In QDA, we assume that

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{2} \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}, \ k=1,2$$

and thus

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\frac{\pi_1/\sigma_1}{\pi_2/\sigma_2} - \frac{1}{2\sigma_1^2}(x-\mu_1)^2 + \frac{1}{2\sigma_2^2}(x-\mu_2)^2$$
$$= c_0 + c_1 x + c_2 x^2.$$

➔ Related to 2^{nd} order polynomial logistic regression.

Comparison of Classification Methods (3/3)

- kNN classification is done by taking

$$\hat{Y}(x) = \arg\max_{k' \in \{1,\ldots,K\}} \frac{1}{k} \sum_{i: x_i \in \mathcal{N}_k(x)} \mathbf{1}_{\{y_i = k'\}} \left(\approx \Pr(Y = k' | X = x) \right).$$

➔ non-parametric

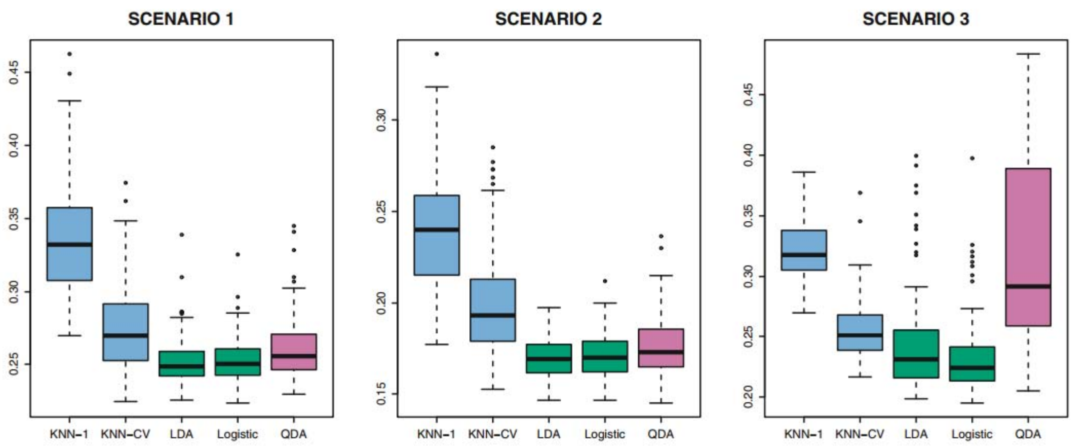
- Flexibility: kNN>QDA>LDA

Example: In the following, we simulate 100 random training sets from several different known distributions with 2 predictors and a binary response variable.

Compare between (1) kNN-1, (2) kNN-CV, (3) Logistic Regression, (4) LDA, (5) QDA.

Example (1/2)

- **Scenario 1:** 20 training observations in each class, uncorrelated normal with different means.
- **Scenario 2:** Same as Scenario 1 but with correlation of -0.5 among the two predictors.
- **Scenario 3:** The two predictors generated from the student t-distribution, with 50 observations per class.



Example (2/2)

- **Scenario 4:** Normal with correlation of 0.5 in class 1 and -0.5 in class 2. Means differ between classes.
- **Scenario 5:** Normal with uncorrelated predictors. Responses were sampled $\Pr(Y=1|X=x) = \frac{e^{\beta_0+\beta_1 X_1^2+\beta_2 X_2^2+\beta_3 X_1 X_2}}{1+e^{\beta_0+\beta_1 X_1^2+\beta_2 X_2^2+\beta_3 X_1 X_2}}$.
- **Scenario 6:** Same as Scenario 5, but responses sampled from a more complicated non-linear function.