STA 314: Statistical Methods for Machine Learning I

Lecture 7 - Discriminant Analsysis: LDA ,QDA, Naive Bayes

Xin Bing

Department of Statistical Sciences
University of Toronto

Review

In the last lecture, we have learned the logistic regression for binary classification with $Y \in \{0, 1\}$.

- Estimating the Bayes rule at any observation $x \in \mathbb{R}^p$ is equivalent to estimate the conditional probability $\mathbb{P}(Y = 1 \mid X = x)$.
- Logistic regression parametrizes the conditional probability by

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{e^{\beta_0 + x^\top \boldsymbol{\beta}}}{1 + e^{\beta_0 + x^\top \boldsymbol{\beta}}}.$$

- We estimate the coefficients by using MLE which can be solved by (stochastic) gradient descent.

Discriminant Analysis

- Logistic regression directly parametrizes $\mathbb{P}(Y = 1 \mid X = x)$.
- By contrast, **Discriminant Analysis** parametrizes the distribution of $X \mid Y = 1$ and $X \mid Y = 0$.
- What does this buy us? By Bayes' theorem,

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1)}{\mathbb{P}(X = x)}.$$

  Thus,

$$\begin{aligned} & \mathbb{P}(Y = 1 \mid X = x) \ \geq \ \mathbb{P}(Y = 0 \mid X = x) \\ \iff \ & \mathbb{P}(X = x \mid Y = 1)\mathbb{P}(Y = 1) \ \geq \ \mathbb{P}(X = x \mid Y = 0)\mathbb{P}(Y = 0) \end{aligned}$$

- The distributions of $X \mid Y = k$, $k \in \{0, 1\}$, are typically assumed to be normal distributions.

Why not Logistic Regression?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Discriminant analysis does not suffer from this problem.

- If n is small and the distribution of $X \mid Y = k$ is correctly specified, discriminant analysis again has better performance than the logistic regression model.

- Discriminant analysis is more suitable for multi-class classification problems.

Notation for discriminant analysis

Suppose we have K classes, $C = \{0, 1, 2, \ldots, K - 1\}$. For any $k \in C$,

- We let

$$\pi_k = \mathbb{P}(Y = k)$$

  be the **prior** probability that a randomly chosen observation comes from the kth class.

- Let

$$f_k(X) = \mathbb{P}(X = x \mid Y = k)$$

  denote the **density function** of $X = x$ from class k.

  In discriminant analysis, parametric assumption is assumed on $f_k(X)$.

The Bayes rule

- By the Bayes' theorem,

$$p_k(x) := \mathbb{P}(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell \in C} \pi_\ell f_\ell(x)}$$

  is called **posterior** probability, i.e. the probability that an observation belongs to the kth class given its feature.

- According to the Bayes classifier, we should classify a new point x according to

$$\arg\max_{k \in C} \; p_k(x) = \arg\max_{k \in C} \; \frac{\pi_k f_k(x)}{\sum_{\ell \in C}^{K} \pi_\ell f_\ell(x)} = \arg\max_{k \in C} \; \pi_k f_k(x).$$

Discriminant Analysis for $p = 1$

- Assume that $X \mid Y = k \sim N(\mu_k, \sigma_k^2)$ is normal for all $k \in C$. Specifically,

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

- **Linear Discriminant Analysis (LDA)** further assumes

$$\sigma_0 = \sigma_1 = \cdots = \sigma_{K-1} = \sigma.$$

- As a result,

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

Linear Discriminant Analysis for $p = 1$

- The Bayes rule classifies $X = x$ to

$$\begin{aligned} \arg\max_{k \in C} \; p_k(x) &= \arg\max_{k \in C} \; \log\left(p_k(x)\right) \\ &= \arg\max_{k \in C} \; \underbrace{\frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k}_{\delta_k(x)} \qquad \text{verify!.} \end{aligned}$$

  The name LDA is due to the fact that the discriminant function, $\delta_k(x)$, is a linear function in x.

- For binary case, i.e. $K = 2$, if the priors are equal $\pi_0 = \pi_1$ and suppose $\mu_1 \geq \mu_0$, then the Bayes classifier assigns $X = x$ to

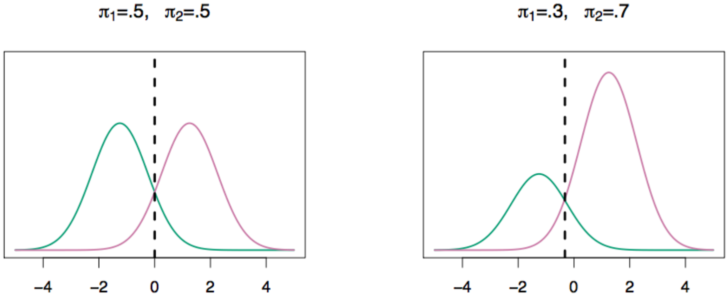
$$\begin{cases} 0 & \text{if } x < \frac{\mu_0 + \mu_1}{2} \\ 1 & \text{if } x \geq \frac{\mu_0 + \mu_1}{2} \end{cases}$$

  The line $x = (\mu_0 + \mu_1)/2$ is called the Bayes decision boundary.

Example of LDA in binary classification

Consider $\mu_0 = -1.5$, $\mu_1 = 1.5$, and $\sigma = 1$. The curves are $p_0(x)$ (green) and $p_1(x)$ (red). The dashed vertical lines are the Bayes decision boundary.

$$f^*(x) = \begin{cases} 0 & \text{if } x < \frac{\mu_0+\mu_1}{2} = 0 \\ 1 & \text{if } x \geq \frac{\mu_0+\mu_1}{2} = 0 \end{cases}$$



Compute the Bayes classifier

- If we know $\mu_0, \ldots, \mu_{K-1}$, σ and $\pi_0, \ldots, \pi_{K-1}$, then we can construct the Bayes rule

$$\arg\max_{k \in C} \delta_k(x) = \arg\max_{k \in C} \left\{ \frac{\mu_k}{\sigma^2} x - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k \right\}.$$

- However, we typically don't know these parameters. We need to use the training data to estimate them!

Estimation under LDA

Given training data $(x_1, y_1), \ldots, (x_n, y_n)$, for all $k \in C$,

- We have
$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}.$$
- We estimate π_k by
$$\hat{\pi}_k = \frac{n_k}{n}.$$
- We estimate μ_k, and $\sigma = 1$ by
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$
$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{u}_k)^2.$$

These are actually the MLE.

The LDA classifier

- We estimate $\delta_k(x)$ by the plug-in estimator

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

- The LDA classifier assigns x to the class with the largest $\hat{\delta}_k(x)$.

- How about the case when $p > 1$?

Linear Discriminant Analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors $X = (X_1, ..., X_p)$.
- Recall that the posterior probability has the form

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell \in C} \pi_\ell f_\ell(x)},$$

- Now, we assume $X \mid Y = k \sim N_p(\mu_k, \Sigma)$, that is,

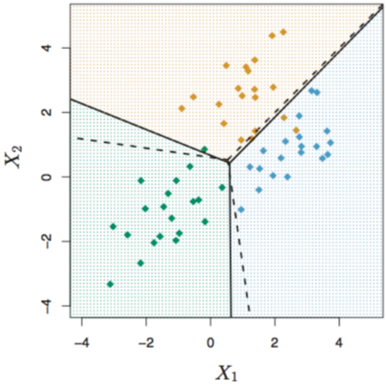
$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}.$$

- Similarly, we assign x to the class with the largest discriminant function

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

- The Bayes decision boundaries are the set of x for which $\delta_k(x) = \delta_\ell(x)$ for $k \neq \ell$, which are again **linear hyperplanes** in x.

Example



There are three classes (orange, green and blue) with two features X_1 and X_2. Dashed lines are the Bayes decision boundaries. Solid lines are their estimates based on the LDA. 2

Estimation under LDA for $p > 1$

The same as before, given training data $(x_1, y_1), \ldots, (x_n, y_n)$, for any $k \in C$,

- We have
$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}.$$
- We estimate π_k by
$$\hat{\pi}_k = \frac{n_k}{n}.$$

The slight difference is to estimate μ_k and Σ by

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{u}_k)(x_i - \hat{u}_k)^\top.$$

A plugin rule for estimating discriminant functions

- We use the plugin estimator

$$\hat{\delta}_k(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \hat{\pi}_k, \quad \forall k \in C.$$

- The resulting LDA classifier is

$$\arg\max_{k \in C} \hat{\delta}_k(x).$$

Logistic Regression versus LDA

For a binary classification problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_0(x)}\right) = c_0 + c_1x_1 + \cdots + c_px_p,$$

which has the same linear form as logistic regression.

- LDA makes more assumption by specifying $X \mid Y$.
- The parameters are estimated differently.
  - Logistic regression uses the conditional likelihood based on $\mathbb{P}(Y|X)$ (known as discriminative learning).
  - LDA uses the full likelihood based on $\mathbb{P}(X, Y)$ (known as generative learning).
- If classes are well-separated, then logistic regression is not advocated.
- Despite these differences, in practice they often perform similarly.

LDA on the Default Data

Classify whether or not an individual will default on the basis of credit card balance and student status. The confusion matrix on default data.

		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

- The training error rate is $(23 + 252)/10000 = 2.75\%$.
- **False positive rate (FPR)**: The fraction of negative examples that are classified as positive: $23/9667 = 0.2\%$ in default data.
- **False negative rate (FNR)**: The fraction of positive examples that are classified as negative: 75.7% in default data.
- For a credit card company that is trying to identify high-risk individuals, an error rate of $252/333 = 75.7\%$ among individuals who default is unacceptable.

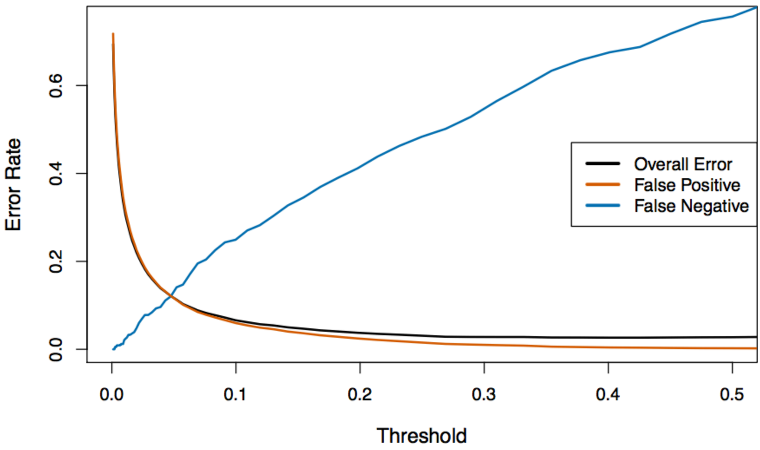
Types of Errors for binary classification

- The false negative rate is too high. How can we modify the LDA rule to lower the FNR?
- The current classifier is based on the rule

$$\mathbb{P}(\text{default} = \textit{yes} \mid X = x) \geq 0.5.$$

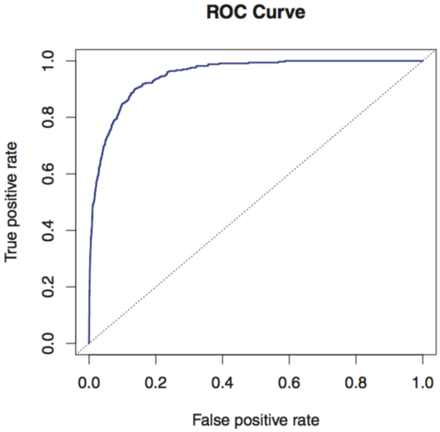
- We can achieve better balance between FPR and FNR by varying the threshold:
  - To lower FNR, we reduce the number of negative predictions. Classify $X = x$ to *yes* if
  $$\mathbb{P}\left(Y = \textit{yes} \mid X = x\right) \geq \text{thresh}.$$
  for some thresh < 0.5.

Trade-off between FPR and FNR



ROC Curve

The ROC curve is a popular graphic for simultaneously displaying FPR and TPR for all possible thresholds.



The overall performance of a classifier, summarized over all thresholds, is given by the area under the curve (AUC). High AUC is good.

More metrics in the binary classification

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

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*	

The above also defines **sensitivity** and **specificity**.

Other forms of Discriminant Analysis

Recall that LDA specifies

$$X \mid Y = k \sim N(\mu_k, \Sigma), \qquad \forall k \in C.$$

Other discriminant analyses change the specifications for $X \mid Y = k$.

- **Quadratic discriminant analysis** (QDA) assumes

  $$X \mid Y = k \sim N(\mu_k, \Sigma_k), \qquad \forall k \in C,$$

  by allowing different Σ_k across all classes.

- **Naive Bayes** assumes

  $$X_1, \ldots, X_p \text{ are independent given } Y = k.$$

  For Gaussian density, this means that Σ_k's are diagonal.

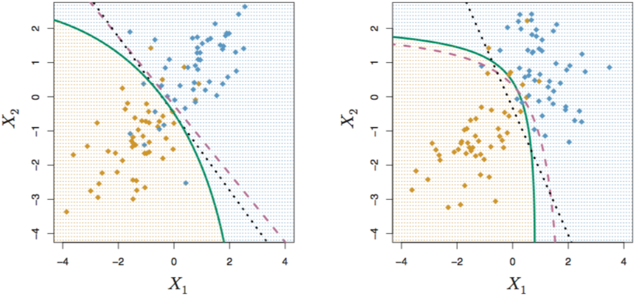
- Many other forms, by proposing specific density models for $X \mid Y = k$, including nonparametric approaches.

Quadratic Discriminant Analysis

In QDA, because of the different Σ_k's, the Bayes classifier assigns an observation $X = x$ to the class for which

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

is largest. So, the decision boundary is quadratic in x.



Decision boundaries of the Bayes classifier (purple dashed), LDA (black dotted), and QDA (green solid) in two scenarios.

Estimation of QDA

Given training data $(x_1, y_1), \ldots, (x_n, y_n)$, for any $k \in C$,

- We have
$$n_k = \sum_{i=1}^{n} 1\{y_i = k\}.$$
- We estimate π_k by
$$\hat{\pi}_k = \frac{n_k}{n}.$$
- We estimate μ_k and Σ by
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$
$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{u}_k)(x_i - \hat{u}_k)^\top.$$
- Plugin estimator for $\delta(x)$.

Potential problems for LDA and QDA in high dimension

- LDA: we have
$$(K-1) + pK + \frac{p(p-1)}{2}$$
number of parameters to estimate.

- QDA: we have
$$(K-1) + pK + \frac{p(p-1)}{2}K$$
number of parameters to estimate.

- The estimation error is large when p is large comparing to n.

Naive Bayes

Naive Bayes assumes that features are *independent* within each class.

Useful when p is large, whence QDA and even LDA break down.

- Under Gaussian distributions, naive Bayes assumes each Σ_k is diagonal. The decision boundary is determined by

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

- It is easy to extend it to mixed features (quantitative and categorical).
- Despite the strong independence assumption, naive Bayes often produces good classification results.