STAT 432: Basics of Statistical Learning

Linear Classification Models

Shiwei Lan, Ph.D. <shiwei@illinois.edu>

http://shiwei.stat.illinois.edu/stat432.html

March 13, 2019

University of Illinois at Urbana-Champaign

Outline

- Classification problems and the Bayes rule
- Logistic Regression
- LDA and QDA

Classification Problems

Classification Problems

- We have n observations, each with a p dimensional covariate X, and a binary outcome Y.
- Training data: $\mathcal{D}_n = \{x_i, y_i\}_{i=1}^n$:
  - $x_i \in \mathbf{R}^p$,
  - $y_i \in \{0, 1\}$ (we could use other codings).
- The goal is to find a (hard) classifier

$$f : \mathbf{R}^p \longrightarrow \{0, 1\}$$

Classification Problems

- Similar to the regression models, the optimal classifier is defined as the one that minimizes the risk, with a loss function

$$\mathsf{R}(f) = \mathsf{E}\left[L\big(Y, f(X)\big)\right]$$

- The most common choice of L for classification problems is the 0–1 loss defined as

$$L\big(a, b\big) = \begin{cases} 0 & \text{if} \quad a = b \\ 1 & \text{if} \quad \text{o.w.} \end{cases}$$

Underlying Model

- We can assume that the underlying model for the probability of $Y = 1$ at each target point x_0 is

$$\eta(x_0) = \mathsf{P}(Y = 1|X = x_0)$$

- What is the best classifier? If we use the 0–1 loss, we try to find f to minimize

$$\mathsf{R}(f) = \eta(x_0)I\big(f(x_0) = 0\big) + \big(1 - \eta(x_0)\big)I\big(f(x_0) = 1\big)$$

- Hence, its better to choose $f(x_0) = 0$ if $\eta(x_0) < 1/2$ and 1 o.w.

Underlying Model

- Apply this to all possible points, we have the optimal rule

$$f_B(x) = \arg\min_f \mathsf{R}(f) = \begin{cases} 1 & \text{if} \quad \eta(x) > 1/2 \\ 0 & \text{if} \quad \eta(x) \leq 1/2. \end{cases}$$

- Note: if $\eta(x) = 1/2$, it doesn't matter what the decision rule is because the loss is 1/2 either way.
- This optimal rule is called the Bayes Rule, and the corresponding risk $\mathsf{R}(f_B)$ is referred to as the Bayes risk or Bayes error.
- We introduce two approaches: logistic regression and discriminate analysis.

Logistic Regression

Motivation

- The logistic regression belongs to a class of generalized linear models (GLM)
- In linear regression, we have $Y = X^{\mathsf{T}}\boldsymbol{\beta} + \epsilon$, however, the outcome is continuous.
- When Y takes only 0 or 1, we can model it using a Bernoulli distribution, i.e.,

$$Y \sim \mathrm{Ber}(p)$$

- To incorporate covariates, we let $p = \eta(x)$, hence, the conditional distribution is

$$Y|X = x \quad \sim \quad \mathrm{Ber}(\eta(x))$$

Motivation

- We can choose many functional forms of $\eta(\cdot)$, however, linear functions seems to be the most convenient and interpretable.
- This motivate us to consider some link function ϕ that transform $\eta(x)$ into $x^\mathsf{T}\boldsymbol{\beta}$

$$\phi(\eta(x)) = x^\mathsf{T}\beta$$

- For logistic regression, we use the logit link function, which is also called the log-odds:

$$\phi(a) = \mathsf{logit}(a) = \log\left(\frac{a}{1-a}\right)$$

Motivation

- From some simple derivations, if we set

$$\text{logit}(\eta(x)) = x^\mathsf{T}\boldsymbol{\beta}$$

  then

$$\begin{aligned} \log \frac{\eta(x)}{1-\eta(x)} &= x^\mathsf{T}\beta \\ \implies \quad \eta(x) &= \frac{\exp(x^\mathsf{T}\beta)}{1+\exp(x^\mathsf{T}\beta)} \end{aligned}$$

- Hence, for logistic regression, the conditional distribution of Y given $X = x$ is

$$\text{Ber}\left(\frac{\exp(x^\mathsf{T}\beta)}{1+\exp(x^\mathsf{T}\beta)}\right)$$

Fitting Logistic Models

- The common technique for solving the $\boldsymbol{\beta}$ parameters is through maximizing the log-likelihood
- Since the response variable Y follows a Bernoulli distribution, the likelihood for a single observation is

$$p(Y = y_i | X = x_i) = \eta(x_i)^{y_i} [1 - \eta(x_i)]^{1-y_i}$$

- Using the logit link function, and a set of observations $\{x_i, y_i\}_{i=1}^{n}$, we have

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \, p(y_i | x_i, \boldsymbol{\beta})$$

Fitting Logistic Models

- This can be simplified into

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log \left\{ \eta(x_i)^{y_i} [1 - \eta(x_i)]^{1-y_i} \right\} \\ &= \sum_{i=1}^{n} y_i \log \frac{\eta(x_i)}{1 - \eta(x_i)} + \log[1 - \eta(x_i)] \\ &= \sum_{i=1}^{n} y_i x_i^{\mathsf{T}} \boldsymbol{\beta} - \log[1 + \exp(x_i^{\mathsf{T}} \boldsymbol{\beta})]\end{aligned}$$

- Hence we solve for $\boldsymbol{\beta}$ that maximizes the log-likelihood

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\arg\max}\ \ell(\boldsymbol{\beta})$$

Newton-Raphson

- To solve for $\widehat{\boldsymbol{\beta}}$, we use Newton's method (second order method)
- Choose an initial value $\boldsymbol{\beta}^0$, then update $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\mathsf{T}}\right]^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

where

$$\begin{aligned} \text{(gradient)} \quad \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n y_i x_i - \sum_{i=1}^n \frac{\exp(x_i^\mathsf{T} \boldsymbol{\beta}) x_i}{1 + \exp(x_i^\mathsf{T} \boldsymbol{\beta})} \\ \text{(Hessian)} \quad \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\mathsf{T}} &= -\sum_{i=1}^n x_i x_i^\mathsf{T} \eta(x_i)[1 - \eta(x_i)] \end{aligned}$$

Penalized Logistic Regression

- When the number of variables is large, we may consider a sparse model. The idea is similar to Lasso or Ridge in linear regressions.
- Ridge penalty

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \quad - 2\ell(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|^2$$

- Lasso penalty

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \quad - 2\ell(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1$$

Revisiting the Bayes Rule

Bayes Rule

- Instead of modeling the conditional probability, we may consider to view the problem in a different way.
- Recall that we simply need the decision rule to be

$$f_B(x) = \begin{cases} 1 & \text{if} \quad \eta(x) > 1/2 \\ 0 & \text{if} \quad \eta(x) \leq 1/2. \end{cases}$$

- Essentially, we just need to compare

$$\mathsf{P}(Y = 1 | X = x) \quad \text{vs.} \quad \mathsf{P}(Y = 0 | X = x)$$

Bayes Rule

- We can apply the Bayes rule:

$$\mathsf{P}(Y=1|X=x) = \frac{\mathsf{P}(X=x|Y=1)\mathsf{P}(Y=1)}{\mathsf{P}(X=x)}$$
$$\mathsf{P}(Y=0|X=x) = \frac{\mathsf{P}(X=x|Y=0)\mathsf{P}(Y=0)}{\mathsf{P}(X=x)}$$

- Here we need a few additional quantities: the priors and the conditional probabilities.
- Noticing that the denominator does not play a role in the comparison.

Bayes Rule

- Let prior probabilities be

$$\pi = \mathsf{P}(Y = 1) \quad \text{and} \quad (1 - \pi) = \mathsf{P}(Y = 0)$$

and define the conditional densities of X as

$$f_1 = \mathsf{P}(X = x | Y = 1) \quad \text{and} \quad f_0 = \mathsf{P}(X = x | Y = 0).$$

- The Bayes rule can also be written as

$$f_B(x) = \begin{cases} 1 & \text{if} \quad \pi f_1(x) > (1 - \pi) f_0(x) \\ 0 & \text{o.w.} \end{cases}$$

Bayes Rule

- The prior probabilities: $\mathsf{P}(Y=0)$ and $\mathsf{P}(Y=1)$
  - Prior knowledge of the likelihood of occurrence for each class
  - May be used to make a decision without any extra knowledge
- The posterior probabilities: $\mathsf{P}(Y|X)$
  - The updated probabilities after observing $X=x$
- Bayes decision rule combines them to achieve the minimum risk

$$f_B(x) = 1 \;\text{ if }\; \frac{f_1(x)}{f_0(x)} > \frac{1-\pi}{\pi}; \;\text{and } 0 \;\text{ o.w.}$$

Bayes Rule

- The decision boundary can be used to describe the optimal rule:

$$\{x : \pi f_1(x) = (1-\pi) f_0(x)\}$$

- Linear methods for classification: the classification rules with $f_B(x)$ being linear in x, or equivalently, classification rules with linear decision boundaries.

Multi-Class Problems

- The Bayes rule can be easily generated to multi-class problems, where $y \in \{1, \dots K\}$. The classifier is

$$f : \mathbf{R}^p \longrightarrow \{1, \dots, K\}$$

- The optimal rule is

$$f_B(x) = \arg\max_k \mathsf{P}(Y = k | X = x) = \arg\max_k \pi_k f_k(x)$$

  where π_k is prior probability and $f_k(x)$ is the conditional density.
- Classify x to the most probable class by comparing $\mathsf{P}(Y | X = x)$, or the product of prior and conditional density.

Binary vs. Multi-Class

- Many binary classifiers can also handle multi-classes, such as discriminate analysis (LDA, QDA, NB), logistic regression, kNN and random forests. But for some others, the extension is non-trivial (e.g. SVM).
- There are some naive (although may not be optimal) ways to apply a binary classifier on a classification problem with $K > 2$ categories.
  - Train K one-vs-other classifiers
  - Train $K(K-1)/2$ pairwise classifiers

  Then we can combine the results to get a consensus prediction.

Discriminant Analysis

A Motivation

- Lets look at a naive approach: “1 vs. others”. The following example is demonstrated by fitting a linear regression for categorical outcomes, mainly for easy understandings (although logistic regression is a better choice).
- For the outcome Y, which may fall into categories $1, \ldots, K$, define a vector of indicators $(Y_1, \ldots, Y_K)$

$$Y_k = 1 \quad \text{if} \quad Y = k$$

  - Each vector $(Y_1, \ldots, Y_K)$ has a single 1.
  - The n training samples form an $n \times K$ indicator response matrix $\mathbf{Y}$, where each row is such an indicator vector.

The Masking Problem

- Fit a linear regression model to each column of $\mathbf{Y}$ simultaneously

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \equiv \mathbf{X}\widehat{\mathbf{B}}$$

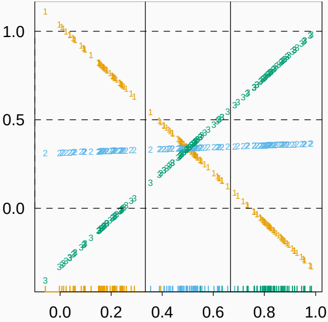
- The k'th column of the parameter matrix $\mathbf{B}$ represents the coefficients for modeling the "likelihood" of being in category k
- Suppose we have a new input x, we can compute $\widehat{f}_k(x) = x^\mathsf{T}\mathbf{B}_{[\,,k]}$, and compare them to predict the class using

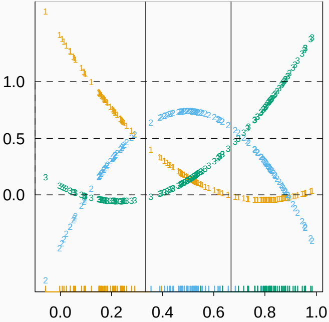
$$Y_{\text{pred}} = \widehat{f}(x) = \underset{k=1,\ldots K}{\arg\max}\ \widehat{f}_k(x)$$

- However, this suffered from some problems:
  - No guarantee that each $\widehat{f}_k(x) \in [0, 1]$
  - Serious masking problem for $K \geq 3$

Masking Problems

Fitting the three-class problem using polynomials





Degree = 1; Error = 0.33

Degree = 2, Error = 0.04

Note: LDA can avoid this problem as it is shown in the previous plot.

Linear Discriminant Analysis

- The idea is to model the distribution of X in each of the classes separately, and then use Bayes theorem to flip things around and obtain P$(Y|X = x)$.
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Naive Bayes (NB)

Bayes Theorem for Classification

- As we demonstrated earlier (Bayes rules), the conditional probability can be formulated using Bayes Theorem:

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

  where $f_k(x)$ is the conditional density function of $X|Y=k$, and $\pi_k = \mathsf{P}(Y=k)$ is the prior probability.

LDA

- The best prediction is picking the one that maximizing the posterior

$$\arg\max_{k}\ \pi_k f_k(x)$$

- Linear discriminate analysis (LDA) assumes that $f_k(x)$ is normal density
- Suppose we model each class density as multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, and assume that the covariance matrices are the same across all k, i.e., $\Sigma_k = \Sigma$. Then the

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x - \boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma^{-1}(x - \boldsymbol{\mu}_k)\right]$$

LDA

- The log-likelihood function for the conditional distribution is

$$\begin{aligned}\log f_k(x) =& -\log((2\pi)^{p/2}|\Sigma|^{1/2}) - \frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k)\\ =& -\frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k) + \text{constant}\end{aligned}$$

- Hence we just need to select the category that attains the highest posterior density (MAP: maximum a posteriori):

$$\begin{aligned}Y_{pred} = \widehat{f}(x) &= \arg\max_k \ \log\big(\pi_k f_k(x)\big)\\ &= \arg\max_k \ -\frac{1}{2}(x-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x-\boldsymbol{\mu}_k) + \log(\pi_k)\end{aligned}$$

Interpretations of LDA

- The term $(x - \boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(x - \boldsymbol{\mu}_k)$ is simply the Mahalanobis distance between x and the centroid $\boldsymbol{\mu}_k$ for class k
- Classify x to the class with the closest (in terms of Mahalanobis distance) centroid, while also adjust for the prior.
- Special case: $\Sigma = \mathbf{I}$ (only Euclidean distance is needed)

$$\arg\max_k \; -\frac{1}{2}\|x - \boldsymbol{\mu}_k\|^2 + \log(\pi_k)$$

Decision Boundary

- Noticing that that quadratic term can be simplified to

$$\begin{aligned}
&-\frac{1}{2}(x-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma^{-1}(x-\boldsymbol{\mu}_k) \\
&= x^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_k + \text{irrelevant constants}
\end{aligned}$$

- Then the discriminant function is defined as

$$\begin{aligned}
\delta_k(x) &= x^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_k + \log \pi_k \\
&= \mathbf{w}_k^{\mathsf{T}}x + b_k,
\end{aligned}$$

- We can calculate $\mathbf{w}_k$'s and b_k's for each class k from the data.

Decision Boundary

- The decision boundary function between class k and l is

$$\begin{aligned} & \mathbf{w}_k^\mathsf{T} x + b_k = \mathbf{w}_l^\mathsf{T} x + b_l \\ \Leftrightarrow \quad & (\mathbf{w}_k - \mathbf{w}_l)^\mathsf{T} x + (b_k - b_l) = 0 \\ \Leftrightarrow \quad & \widetilde{\mathbf{w}}^\mathsf{T} x + \widetilde{b} = 0 \end{aligned}$$

- Since $\mathbf{w}_k = \Sigma^{-1} \boldsymbol{\mu}_k$ and $\mathbf{w}_l = \Sigma^{-1} \boldsymbol{\mu}_l$, the decision boundary has the directional vector

$$\widetilde{\mathbf{w}} = \Sigma^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)$$

Interpretations of LDA

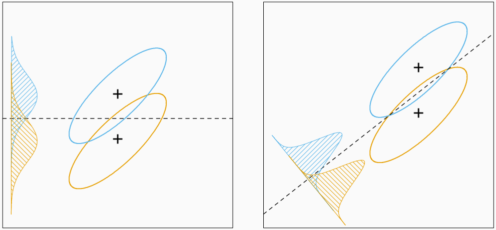


FIGURE 4.9. *Although the line joining the centroids defines the direction of greatest centroid spread, the projected data overlap because of the covariance (left panel). The discriminant direction minimizes this overlap for Gaussian data (right panel).*

Parameter Estimations in LDA

- We estimate the LDA parameters from the training data
    - Prior probabilities: $\widehat{\pi}_k = n_k/n = n^{-1}\sum_k \mathbf{1}\{y_i = k\}$, where n_k is the number of observations in class k.
    - Centroid: $\widehat{\boldsymbol{\mu}}_k = n_k^{-1}\sum_{i:\, y_i=k} x_i$
    - Pooled covariance (since we assume that covariance matrices are the same across different classes):

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

Quadratic Discriminant Analysis

- Quadratic Discriminant Analysis (QDA) simply abandons the the common covariance matrix assumption. Hence, the Σ_k's are not equal.
- In this case, the determinant $|\Sigma_k|$ of each covariance matrix will be different. The MAP decision becomes

$$\begin{aligned} & \max_k \ \log\big(\pi_k f_k(x)\big) \\ =& \max_k \ -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma_k^{-1}(x-\boldsymbol{\mu}_k) + \log(\pi_k) \\ =& x^{\mathsf{T}}\mathbf{W}_k x + \mathbf{w}_k^{\mathsf{T}} x + b_k \end{aligned}$$

- A quadratic decision boundary between class k and l

$$\{x : x^{\mathsf{T}}(\mathbf{W}_k - \mathbf{W}_l)x + (\mathbf{w}_k^{\mathsf{T}} - \mathbf{w}_l^{\mathsf{T}})^{\mathsf{T}} x + (b_k - b_l) = 0\}$$

Estimations in QDA

- We estimate the QDA parameters from the training data
  - Prior probabilities: $\widehat{\pi}_k = n_k/n = n^{-1}\sum_k \mathbf{1}\{y_i = k\}$, where n_k is the number of observations in class k.
  - Centroid: $\widehat{\boldsymbol{\mu}}_k = n_k^{-1}\sum_{i:\, y_i=k} x_i$
  - Sample covariance matrix for each class:

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

LDA vs. QDA

- More parameters in QDA than LDA, especially when p is large
- Both are extremely simple to implement and perform well on real classification problems
- We can include selected quadratic terms of the covariates, such as X_1X_2 or X_1^2, and still perform LDA

LDA vs. QDA

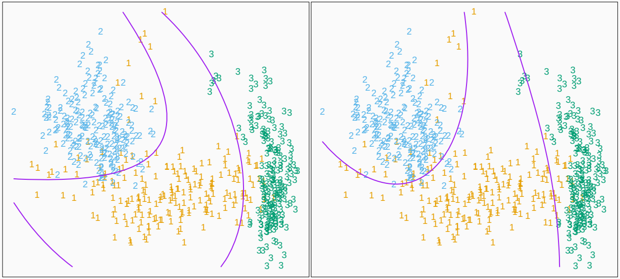


FIGURE 4.6. *Two methods for fitting quadratic boundaries. The left plot shows the quadratic decision boundaries for the data in Figure 4.1 (obtained using LDA in the five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$). The right plot shows the quadratic decision boundaries found by QDA. The differences are small, as is usually the case.*

Reduced Rank LDA

- Low-dimensional structure of the data
  - The K centroids $(\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K)$ in p-dimensional input space span a subspace of rank $K - 1$, denote this subspace as H
  - For any point x, we can project it onto H, and make a comparison in this reduced space
  - Think of these K centroids (their differences) as basis of the space.
  - When $K \ll p$, this is a considerable drop in dimension
  - When $K = 3$, we can view the data in a two dimensional plot
  - When K is large, a natural way is to perform PCA on these K mean vectors (treating each centroid as an observation). They are usually called discriminant coordinates or canonical variates.

Discriminant Coordinates

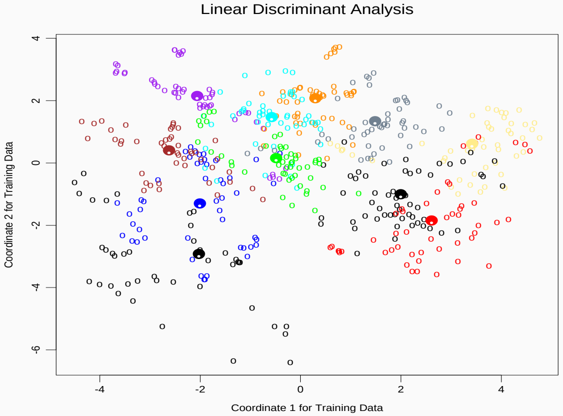


FIGURE 4.4. *A two-dimensional plot of the vowel training data. There are eleven classes with $X \in \mathbb{R}^{10}$, and this is the best view in terms of a LDA model (Section 4.3.3). The heavy circles are the projected mean vectors for each class. The class overlap is considerable.*

Discriminant Coordinates

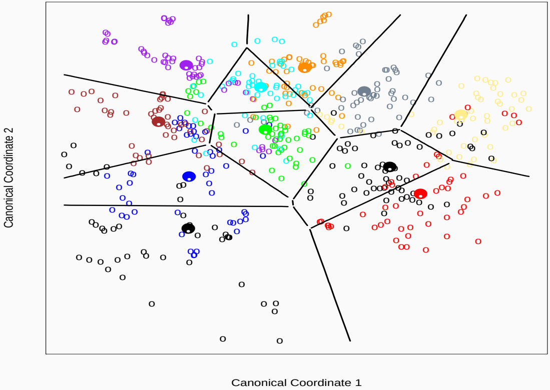


FIGURE 4.11. *Decision boundaries for the vowel training data, in the two-dimensional subspace spanned by the first two canonical variates. Note that in any higher-dimensional subspace, the decision boundaries are higher-dimensional affine planes, and could not be represented as lines.*

Discriminant Analysis in Large p Problems

- When p is large, QDA/LDA may not be applicable, because the inverse of $\widehat{\Sigma}$ may not exist
- Using generalized inverse matrix can easily overfit the data
- A warning sign: Classes are well-separated on the training data could be meaningless for high-dimensional data
- Regularization: sparse LDA, Naive Bayes, RDA

Regularized Discriminant Analysis (RDA)

- Friedman (1989): shrink the separate covariances of QDA toward a common covariance in LDA. Regularized covariance matrices are

$$\widehat{\Sigma}_k(\alpha) = \alpha\widehat{\Sigma}_k + (1-\alpha)\widehat{\Sigma}$$

- $\alpha \in [0,1]$, a continuum of models between LDA and QDA, if $\widehat{\Sigma}$ is the pooled covariance matrix used in LDA
- In practice, chose α using CV.
- We can further shrink Σ_k towards the diagonal covariance, with $\gamma \in [0,1]$

$$\widehat{\Sigma}_k(\alpha,\gamma) = \alpha\widehat{\Sigma}_k + (1-\alpha)\gamma\widehat{\Sigma} + (1-\alpha)(1-\gamma)\widehat{\sigma}^2\mathbf{I}$$

Naive Bayes

- Recall that the optimal decision rule is

$$\arg\max_k \mathsf{P}(Y = k|X = x) = \arg\max_k \pi_k f_k(x)$$

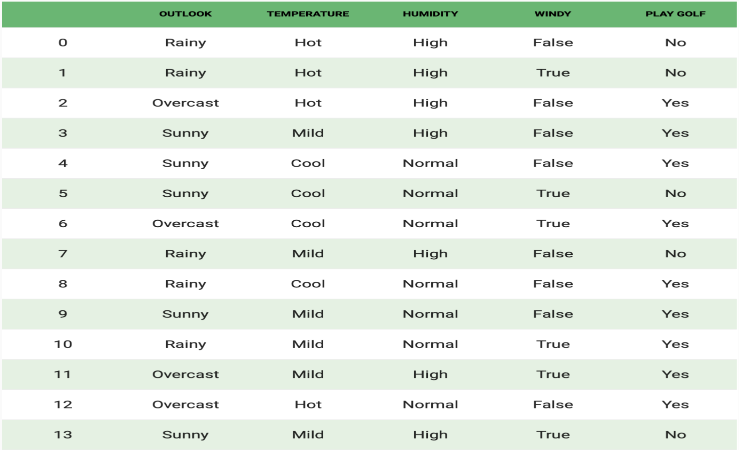
- We can approximate $f_k(x)$ by

$$f_k(x) \approx \prod_{j=1}^{p} f_{kj}(x_j),$$

  meaning that each dimension of x is approximately independently
- $f_{kj}(x_j)$ can be estimated using histograms (discrete), or kernel densities (continuous)
- When Gaussian kernel is used, what is the connection to LDA?

Naive Bayes: golf example

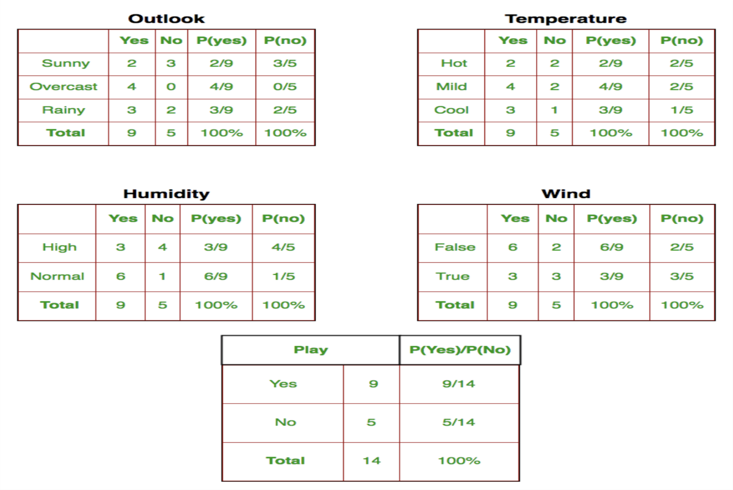
- Consider the weather conditions for playing a game of golf.
- Task: decide whether to play golf depending on given weather condition.



	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naive Bayes: golf example

- Compute the prior and conditional probabilities.



Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

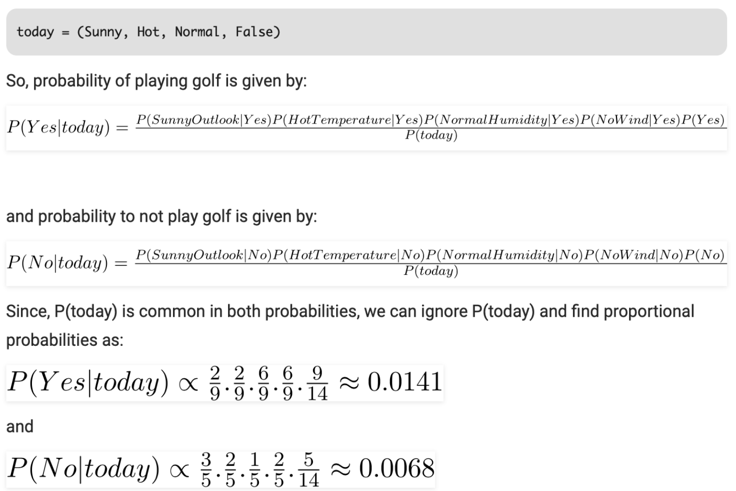
Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

Naive Bayes: golf example

- Make decision (classification) based on covariates (weather conditions).



```
today = (Sunny, Hot, Normal, False)
```

So, probability of playing golf is given by:

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

and probability to not play golf is given by:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

Since, P(today) is common in both probabilities, we can ignore P(today) and find proportional probabilities as:

$$P(Yes|today) \propto \frac{2}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.0141$$

and

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

Logistic Regression vs. LDA

- For LDA, the log-posterior odds between class 1 and 0 are linear in x

$$\log \frac{\mathsf{P}(Y=1|X=x)}{\mathsf{P}(Y=0|X=x)} = \log \frac{\pi_1}{\pi_0} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_0^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\mu}_0 + x^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) = \alpha_0 + x^{\mathsf{T}}\boldsymbol{\alpha}$$

- Logistic model has linear logics by construction

$$\log \frac{\mathsf{P}(Y=1|X=x)}{\mathsf{P}(Y=0|X=x)} = \beta_0 + x^{\mathsf{T}}\boldsymbol{\beta}$$

- Are they the same estimators?

Logistic Regression vs. LDA

- For LDA, the The linearity is a consequence of the Gaussian assumption for the class densities, and the assumption of a common covariance matrix.
- For logistic regression, the linearity comes by construction.
- The difference lies in how the coefficients are estimated.
- Which is more general?
  - LDA assumes Gaussian distribution of X; while logistic leaves the density of X arbitrary
- Logistic model is more general

R Functions

- LDA and QDA: R package MASS, functions lda, qda.
- naive Bayes: R package e1071, function naiveBayes.
- Logistic: R function glm
- General optimization: R function optim