Lecture 8: Linear Discriminant Analysis

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics, University of British Columbia

STAT 447B: Methods for Statistical Learning

September–December 2014

Today's Learning Goals

1. Review of Logistic Regression and Cross Validation
2. Optimal Classification
3. Plotting $\hat{p}(\mathbf{x})$ for 2-Dimensional $\mathbf{x}$
4. Linear Discriminant Analysis (LDA)

Vaso Constriction: Summary

- We compared 2 models
  - `Y ~ Volume`, i.e., $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i$
  - `Y ~ Volume + Rate`, i.e., $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i + \beta_2 r_i$
- 3 comparisons say the second model is better
  - $H_0 : \beta_2 = 0$ is rejected using a test based on approximate normality
  - A likelihood ratio test or equivalent analysis of deviance rejects $H_0 : \beta_2 = 0$
  - The model with Volume and Rate has smaller misclassification rate under cross validation

Digit Recognition

- Again from the UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/machine-learning-databases/mfeat/mfeat-pix`
- 10 classes, one for each of the digits $0, \ldots, 9$
- Can turn this into a 2-class problem by considering only two digits, e.g., “8” and ”9”
- 240 explanatory variables from 15×16 averages of pixels from a grey-scale image of a handwritten digit, taking values 0–7
- Database has 200 cases for each of the 10 digits (“0” data first, then “1” data, etc.)
- We will not compare models yet, just assess the model with linear predictor using all 240 explanatory variables

$$\eta(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{240} x_{i,240}$$

Digit Recognition: Misclassification Rate on Training Data

True y	$\hat{y} = 0$	$\hat{y} = 1$
0 (“8”)	200	0
1 (“9”)	0	200
Misclass. rate	$(0+0)/400 = 0$	

Perfect prediction!

Digit Recognition: Cross-Validated Misclassification Rate

$\hat{y}$ here is from 10-fold cross-validation

True y	$\hat{y} = 0$	$\hat{y} = 1$
0 (“8”)	196	4
1 (“9”)	4	196
Misclass. rate	$(4+4)/400 = 0.02$	

2% error rate

Digit Recognition: How Much Computing Time?

- 400 observations
- 240 explanatory variables
- 241 parameters to estimate $(\beta_0, \ldots, \beta_{240})$
- The logistic regression model is fit 10 times under 10-fold cross-validation
- There is no closed form solution for the maximum likelihood fit. It has to be done numerically by an iterative algorithm.

Clicker questions 1 and 2.

Classification

- Data $(y_1, \mathbf{x}_1)$, $(y_2, \mathbf{x}_2)$, ..., $(y_n, \mathbf{x}_n)$
- $y = 0/1$ codes 2 classes (for now)
- The following argument applies to any classifier, but consider logistic regression
  - Linear predictor $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$
  - We model the probability $\Pr(Y = 1|\mathbf{x})$ as

$$\Pr(Y = 1 \,|\, \mathbf{x}) = \frac{\exp(\eta(\mathbf{x}_i))}{1 + \exp(\eta(\mathbf{x}_i))}$$

- $\Pr(Y = 0|\mathbf{x})$?
  - In general $\Pr(Y = 0|\mathbf{x}) = 1 - \Pr(Y = 1|\mathbf{x})$
  - Logistic regression

$$\Pr(Y = 0 \,|\, \mathbf{x}) = \frac{1}{1 + \exp(\eta(\mathbf{x}_i))}.$$

From Prediction to Classification

- We can estimate $\beta_0, \beta_1, \beta_2, \ldots$ using MLE
- Function `glm` in `R`
- Given $\mathbf{x}$ (new test point)
  - Predict $p(\mathbf{x}) = \Pr(Y = 1 \mid \mathbf{x})$ using the `predict` function in `R`
  - Gives prediction $\hat{p}(\mathbf{x})$
  - Hence predict / classify the unknown class $y(\mathbf{x})$ as
    - 1 if $\hat{p}(\mathbf{x}) \geq 0.5$
    - 0 otherwise

Optimal Classification?

- Is there a better way of going from the prediction $\hat{p}(\mathbf{x})$ to the classification $y(\mathbf{x})$?
- What would the "optimal" rule be?

Misclassification Error is 0/1 Loss

- We have a true value $y = y(\mathbf{x})$ and a prediction $\hat{y} = \hat{y}(\mathbf{x})$
- 0/1 loss function (applies to any number of classes, K)

$$L(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \text{ (no error)} \\ 1 & \text{if } y \neq \hat{y} \text{ (error)} \end{cases}$$

- Find a function (classifier) $\hat{y}(\mathbf{x})$ with smallest expected loss

$$E_{(Y(\mathbf{x}))}\left[L\left(Y(\mathbf{x}), \hat{y}(\mathbf{x})\right)\right] = \min_h E_{(Y(\mathbf{x}))}\left[L\left(Y(\mathbf{x}), h(\mathbf{x})\right)\right]$$

Minimal expected loss = Minimal expected misclassification error

Expected Loss

- Find a function (classifier) $\hat{y}(\mathbf{x})$ such that

$$E_{Y(\mathbf{x})}\left[L\left(Y(\mathbf{x}), \hat{y}(\mathbf{x})\right)\right] \;\leq\; E_{Y(\mathbf{x})}\left[L\left(Y(\mathbf{x}), h(\mathbf{x})\right)\right]$$

  for any other function h

- The expected loss is

$$E_{Y(\mathbf{x})}\left[L\left(Y(\mathbf{x}), \hat{y}(\mathbf{x})\right)\right] = \sum_{k=1}^{K} L\left(c_k, \hat{y}(\mathbf{x})\right) \Pr\left(Y(\mathbf{x}) = c_k\right)$$

  where the c_k code the classes (e.g., 0 and 1 for 2 classes).

The Winner is the Class With the Largest Probability

- As L is 0/1

$$\begin{aligned}\sum_{k=1}^{K} L(c_k, \hat{y}(\mathbf{x}))\, \Pr(Y(\mathbf{x}) = c_k) &= \sum_{c_k \neq \hat{y}(\mathbf{x})} \Pr(Y(\mathbf{x}) = c_k) \\ &= 1 - \Pr(Y(\mathbf{x}) = \hat{y}(\mathbf{x}))\end{aligned}$$

- i.e., the optimal classifier $\hat{y}$ should maximize

$$\Pr(Y(\mathbf{x}) = \hat{y}(\mathbf{x}))$$

- Hence $\hat{y}(\mathbf{x})$ should be the class with the highest (estimated) probability.
- For 2 classes $\hat{y}(\mathbf{x})$ is the class with $\hat{p}(\mathbf{x}) \geq 0.5$.

Optimal?

- The above argument assumes all types of errors have the same magnitude of loss ($L = 1$)
- e.g., with 2 classes there are two types of errors
  - True $y = 0$ but $\hat{y} = 1$
  - True $y = 1$ but $\hat{y} = 0$
  - May have different losses (costs)
- The argument also assumes the prediction model giving $\hat{p}$ (e.g., logistic) is fixed. There may be better prediction models.

Flexible Logistic Regression

- More flexible models: splines, penalized splines, etc.
- e.g., Generalized additive model (GAM)
  - More flexible linear predictor

$$\eta(\mathbf{x}_i) = \beta_0 + \beta_1 f_1(x_{i1}) + \beta_2 f_2(x_{i2}) + \cdots$$

  - Then apply the logistic transformation as before

$$\Pr(Y = 1 \mid \mathbf{x}) = \frac{\exp(\eta(\mathbf{x}_i))}{1 + \exp(\eta(\mathbf{x}_i))}$$

  - Can be done with `gam` in `R`
  - e.g., Vaso constriction data:

```
vaso.gam <- gam(Y ~ s(Volume) + s(Rate),
   data = vaso, family = 'binomial')
```

Vaso Constriction: 10-Fold Cross Validation

- $\hat{p}$ and $\hat{y}$ are from 10-fold cross-validation
- Try $\hat{p}$ from GLM and from GAM
- Misclassification rates

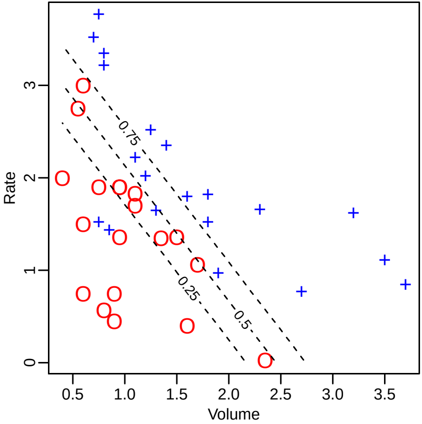
	`glm`		`gam`	
True y	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$
0	14	5	13	6
1	4	16	4	16
Misclass. rate	$(5 + 4)/39 = 0.23$		$(6 + 4)/39 = 0.25$	

- No evidence of improvement from `gam` here

Plotting $\hat{p}(\mathbf{x})$

- Want to visualize the fitted model, say from logistic regression (`gam`)
- Get predictions from the `predict` function for a grid of $\mathbf{x}$ values
- For 2-dimensional $\mathbf{x}$ can plot the predictions against $\mathbf{x}$
  - Use `contour` in `R`
- e.g., for Vaso Constriction Data ...

Vaso Constriction Data: Logistic Fit



Clicker question 3.

Classification by Modelling Y or $\mathbf{X}$?

- So far the statistical model treats the class variable Y as random and the explanatory variables $\mathbf{x}$ as non-random
- e.g.,

$$\Pr(Y = 1 \mid \mathbf{x}) = \frac{\exp(\eta(\mathbf{x}_i))}{1 + \exp(\eta(\mathbf{x}_i))}$$

- Here we are conditioning on $\mathbf{x}$ values, which are non-random, even if they were generated from random variables $\mathbf{X}$
- What about treating $\mathbf{X}$ as random conditional on the class y?

A Model for **X** Conditional on the Class

- Model the distribution of the explanatory variables (features) conditional on each class

$$f(\mathbf{X} \mid Y = c_k) = f_k(\mathbf{X}) \qquad k = 1, \ldots, K$$

(The classes are coded by c_k, e.g., 0, 1 for $K = 2$ classes)

- With prior probabilities $p_k = \Pr(Y = c_k)$, by Bayes' Theorem

$$\Pr(Y = c_k \mid \mathbf{X}) = \frac{f(\mathbf{X} \mid Y = c_k)\, p_k}{f(\mathbf{X})} = \frac{f_k(\mathbf{X})\, p_k}{f(\mathbf{X})} \propto f_k(\mathbf{X})\, p_k$$

Optimal classifier is therefore

$$\hat{y}(\mathbf{X}) = \arg \max_{1 \leq k \leq K} f_k(\mathbf{X})\, p_k$$

Normal Model for $\mathbf{X}|Y = c$

- For example, we can assume that

$$\mathbf{X} \mid Y = c_k \sim \text{MN}\,(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

  (MN = multivariate normal, with dimension the number of variables in **X**)
- The classes differ in their **X** mean vectors
- The class distributions are estimated by

$$\hat{f}_k\,(\mathbf{X}) \sim \text{MN}\left(\hat{\boldsymbol{\mu}}_k, \widehat{\boldsymbol{\Sigma}}\right)$$

  using the sample mean of each group and the pooled sample covariance matrix
- We can then find, for a given **x**, the class k that has the largest $\hat{f}_k(\mathbf{x})\, p_k$

Fisher's Linear Discriminant Analysis for **NORMAL** Populations

Writing f_1 for MN$\,(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and f_2 for MN$\,(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ then

$$f_1(\mathbf{x})\, p_1 \;>\; f_2(\mathbf{x})\, p_2 \quad \Leftrightarrow \quad \log\left(\frac{f_1(\mathbf{x})\, p_1}{f_2(\mathbf{x})\, p_2}\right) > 0 \quad \Leftrightarrow \quad \mathbf{a}^T\mathbf{x} + b \;>\; 0$$

for some vector $\mathbf{a} \in \mathbb{R}^p$ and number $b \in \mathbb{R}$. In other words, boundaries between classes are **linear.** Furthermore, we can estimate this linear boundary because

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}\, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and

$$b = -\frac{1}{2}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\, \boldsymbol{\Sigma}^{-1}\,(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\left(\frac{p_2}{p_1}\right)$$

Classification rule for **NORMAL** populations

We can also write this in term of class probabilities

$$\frac{\Pr(Y = c_1 \,|\, \mathbf{X})}{\Pr(Y = c_2 \,|\, \mathbf{X})} > 1 \quad \Leftrightarrow \quad f_1(\mathbf{x})\, p_1 \;>\; f_2(\mathbf{x})\, p_2$$

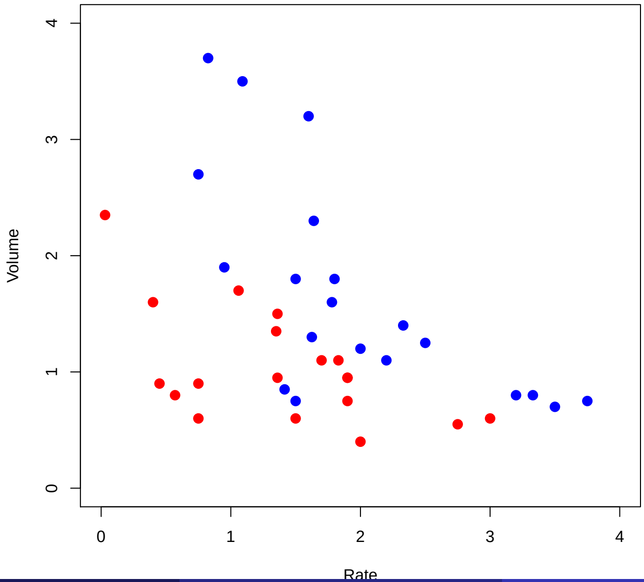
$$\Leftrightarrow \quad \log\left(\frac{f_1(\mathbf{x})\, p_1}{f_2(\mathbf{x})\, p_2}\right) > 0 \quad \Leftrightarrow \quad \mathbf{a}^T\mathbf{x} + b \;>\; 0$$

In fact, for normally distributed features we have

$$\log\left(\frac{\Pr(Y = c_1 \,|\, \mathbf{X})}{\Pr(Y = c_2 \,|\, \mathbf{X})}\right) \;=\; \log\left(\frac{\Pr(Y = c_1 \,|\, \mathbf{X})}{1 - \Pr(Y = c_1 \,|\, \mathbf{X})}\right) \;=\; \mathbf{a}^T\mathbf{x} + b$$

With two classes, we also estimated $\mathbf{a}$ and b using logistic regression.

Vaso Constriction Data



Vaso Constriction: LDA

- First assume that `Volume` and `Rate` are distributed multivariate (bivariate) normal in each class
- Then, the optimal classifier classifies a point $\mathbf{x} = (\text{Volume}, \text{Rate})^T$ in class 1 (red) if

$$\mathbf{a}^T \mathbf{x} + b \; > \; 0$$

where

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

and

$$b = -\frac{1}{2} \, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log \left(\frac{p_2}{p_1} \right)$$

- Furthermore, we can estimate $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ (and even p_1 and p_2) using the sample (How?)

Vaso Constriction: LDA Fit

- We get $\hat{\mathbf{a}} = (-2.77, -2.37)^T$ and $\hat{b} = 7.72$
- Then, the estimated optimal classifier (assuming normality of the features) classifies a point $\mathbf{x} = (\text{Volume}, \text{Rate})^T$ in class 1 (red) if

$$-2.77\,\text{Volume} - 2.37\,\text{Rate} + 7.72 > 0$$

- Furthermore

$$\begin{aligned} &\widehat{P}\,(Y = 1 \mid (\text{Volume}, \text{Rate})) \\ = \; &\frac{\exp\left(-2.77\,\text{Volume} - 2.37\,\text{Rate} + 7.72\right)}{1 + \exp\left(-2.77\,\text{Volume} - 2.37\,\text{Rate} + 7.72\right)} \end{aligned}$$

Vaso Constriction: Plotting the Fit

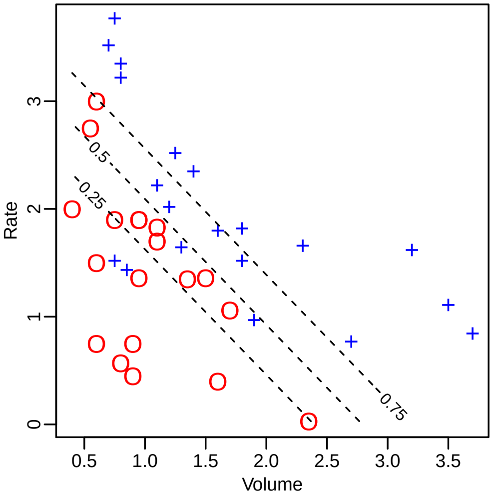
- Now, create a fine grid of `Volume` and `Rate` values, and use the previous formulas to predict

$$\Pr(Y = 1 \mid (\text{Volume}, \text{Rate}))$$

- Plot these posterior probabilities

- We can do this by hand, or using the function `lda` in package `MASS` and its `predict` method

Vaso Constriction Data: LDA Fit



Clicker question 4.

Vaso Constriction Data: Logistic Fit