Ch 4.4.1: Linear Discriminant Analysis

Lecture 8 - CMSE 381

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

February 5, 2024

Announcements

Last time:

- Logistic Regression

Announcements:

- Third homework due Friday! Covers:
  - Mon 2/12 Review Midterm 1
  - Weds 2/14 No class
  - Fri 2/16 Midterm 1
- Office hours
  - Mon: 4-6pm; Tue: 12:30-2:30pm; Wed: 7-9pm; Thu: 12:30-2:30pm

Covered in this lecture

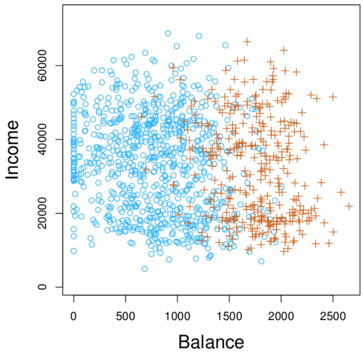
- Bayes theorem
- Linear Discriminant Analysis,
- Quadratic Discriminant Analysis

Section 1

Logistic Regression Review

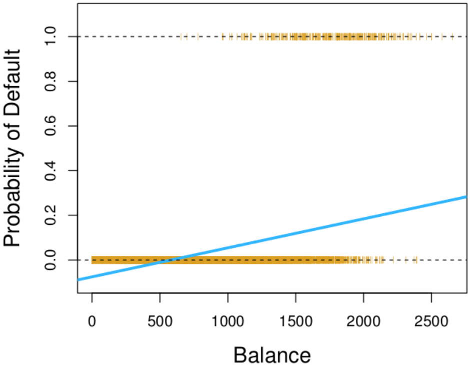
What is classification

- Classification: When the response variable is qualitative
- Goal: Model the probability that Y belongs to a particular category
- Example data:
  $p(\texttt{balance}) = \Pr(\texttt{default} = \texttt{yes} \mid \texttt{balance})$

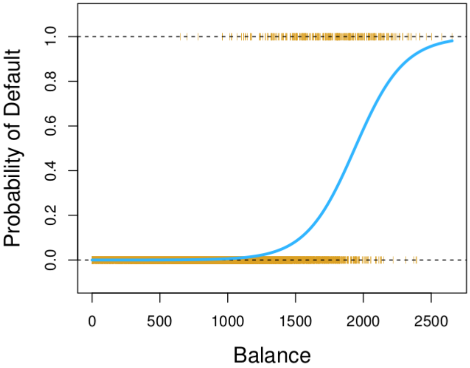


Logistic Regression

$$\Pr(\texttt{default} = \texttt{yes} \mid \texttt{balance}) = \frac{e^{\beta_0 + \beta_1 \texttt{balance}}}{1 + e^{\beta_0 + \beta_1 \texttt{balance}}}$$



Linear Regression



Logistic Regression

Odds

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

- Logistic function is chosen so that odds are linear
- Can take any value from 0 (low odds) to ∞ (high odds)

- 1 in 4 people with odds of $1/3$ will default since $p(X) = 0.25$ and $0.25/(1-0.25) = 1/3$
- 9 in 10 people with odds of 9 will default since $p(X) = 0.9$ and $0.9/(1-0.9) = 9$

How to get logistic function

Assume the (natural) log odds (logits) follow a linear model

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

Solve for $p(x)$:

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

Playing with the logistic function: `https://www.desmos.com/calculator/jzsakksqcm`

Interpreting the coefficients

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

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

β_1 means increasing x by one unit increases the log odds by 1 unit

Estimating Coefficients: Maximum Likelihood Estimation

- **Likelihood:** Probability that data is generated from a model

$$\ell(model) = \Pr[data \mid model]$$

- Find the most likely model

$$\max_{model} \ell(model)$$

- Hard to maximize likelihood, instead maximize log

$$\max_{model} \log(\ell(model))$$

- Strictly increasing log function doesn't change maximum

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

$$\ell(\beta_0, \beta_1) = \prod_{i \mid y_i=1} p(x_i) \prod_{i' \mid y_{i'}=0} (1 - p(x_{i'}))$$

Multiple Logistic Regression

Multiple features:

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

Equivalent to:

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

Multinomial Logistic Regression

What if we have a categorical variable with more than two levels (let's say K of them)?

Plan A

Play the dummy variable game:

Make K the baseline:

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Calculated so that log odds between two classes is linear:

$$\log\left(\frac{\Pr(Y = k \mid X = x)}{\Pr(Y = K \mid X = x)}\right) = \beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p$$

Example

Predict
$Y \in \{\texttt{stroke, overdose, seizure}\}$ for
hospital visits based on Xp

$$\Pr(Y = \texttt{stroke} \mid X = x) = \frac{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x)}{1 + \exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}$$
$$\Pr(Y = \texttt{overdose} \mid X = x) = \frac{\exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}{1 + \exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}$$
$$\Pr(Y = \texttt{seizure} \mid X = x) = \frac{1}{1 + \exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}$$

Plan B: Softmax coding

Treat all levels symmetrically

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}.$$

Calculated so that log odds between two classes is linear

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p.$$

Softmax example

$$\Pr(Y = \texttt{stroke} \mid X = x)$$
$$= \frac{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x)}{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x) + \exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}$$

$$\Pr(Y = \texttt{overdose} \mid X = x)$$
$$= \frac{\exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x)}{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x) + \exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}$$

$$\Pr(Y = \texttt{seizure} \mid X = x)$$
$$= \frac{\exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}{\exp(\beta_{\texttt{str},0} + \beta_{\texttt{str},1}x) + \exp(\beta_{\texttt{OD},0} + \beta_{\texttt{OD},1}x) + \exp(\beta_{\texttt{seiz},0} + \beta_{\texttt{seiz},1}x)}$$

Section 2

Generative Models

Goal:

Another way to approximate
$Pr(Y = k \mid X = x)$

How?
BAYES!!!!

Bayes Theorem

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

- $P(A \mid B)$ (Posterior): probability of A being true given B
- $P(A)$ (Prior): probability of A being true
- $P(B)$ (Marginalization): probability of B being true
- $P(B \mid A)$ (Likelihood): probability of B true given that A is true

Example: Favorite language by year

Example:

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$
\begin{aligned}
P(\text{Fire}|\text{Smoke}) &= \frac{P(\text{Fire})\ P(\text{Smoke}|\text{Fire})}{P(\text{Smoke})} \\
&= \frac{1\% \text{ x } 90\%}{10\%} \\
&= 9\%
\end{aligned}
$$

Example: Favorite language by year

	Fresh	Soph	Junior	Senior	
Python	9	14	13	17	53
R	14	15	10	8	47
	23	29	23	25	

$$P(Y = \texttt{py} \mid X = \texttt{jr}) = \frac{P(Y = \texttt{py}) \cdot P(X = \texttt{jr} \mid Y = \texttt{py})}{P(X = \texttt{jr})}$$

An equivalent formula

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

$$\Leftrightarrow P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(A)P(B|A) + P(A^c)P(B|A^c)}$$

Following book notation

- Classify an observation into one of $K \geq 2$ classes
- π_k is the *prior* probability that a randomly chosen observation comes from the kth class, $P(Y = k)$
- $f_k(X) = \Pr(X \mid Y = k)$ is the density function of X for an observation from the kth class
  - Large $f_k(x)$ if there is high probability that observation in the kth class has $X \approx x$
  - Small if unlikely that an observation in the kth class has $X \approx x$

Bayes to the rescue!

Posterior probability that an observation $X = x$ belongs to the kth class:

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)}$$

- Second equation is Bayes, $p_k(x)$ is our notation for it
- Plug in estimates for π_k and $f_k(x)$ to get an estimate
- Estimate for π_k is easy with a random sample, just take proportion that are kth class
- Estimating desnity f_k is hard(er).....

Section 3

Linear Discriminant Analysis for $p = 1$

Assumptions

Assume $f_k(x)$ is normal/Gaussian:

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

- μ_k = mean of kth class
- σ_k^2 = variance of kth class
- Assume $\sigma_1^2 = \cdots = \sigma_K^2$

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

Bayes Classifier

Same Bayes person, different Bayes definition

Bayes classifier

Assign the class k for which $p_k(x)$ is largest

Finding largest k

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

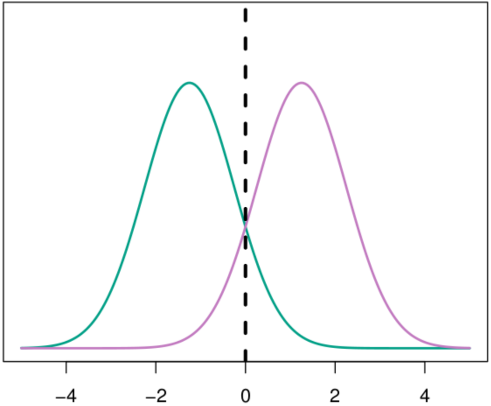
is the same as finding largest k for

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

Example when $K = 2$, $\pi_1 = \pi_2$

Decision boundary:

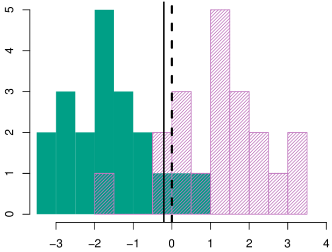
$$x = \frac{\mu_1 + \mu_2}{2}$$



New plan: Linear Discriminant Analysis (LDA)

Estimate

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



- $\hat{\mu}_k = \frac{1}{n_k} \sum_{i|y_i=k} x_i$
- $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i|y_i=k} (x_i - \hat{\mu}_k)^2$
- $\hat{\pi}_k = n_k / n$
- Black solid line: calculated boundary for assignment
- This example, $n_1 = n_2 = 20$, so $\hat{\pi}_1 = \hat{\pi}_2$, so decision bdry half way between sample means
- Optimal Bayes decision boundary dashed line

Example 1

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent proft. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage proft was $X = 4$ last year.

Example 2

Assume the probability that a person defaulted on credit card payment is 10%. Based on the given use LDA to predict the default status of a person with a credit card balance of 1800.

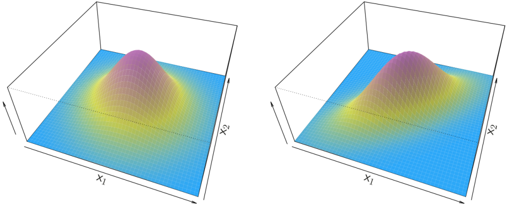
Balance	**Prediction**
0	*No*
500	*No*
1000	*Yes*
1500	*No*
2000	*Yes*
2500	*Yes*

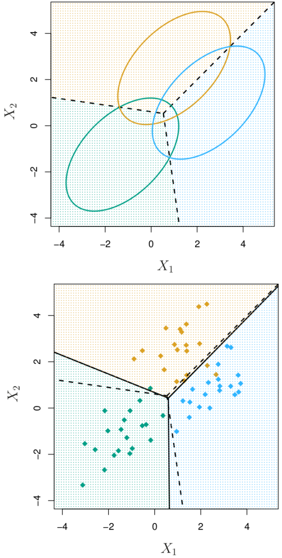
LDA review

- Assume observations in each class come from normal
- Class specific means
- Common variance
- Plug in estimates into Bayes classifier

High dimensional LDA - $p > 1$

What if you have $p > 1$?



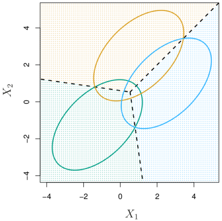


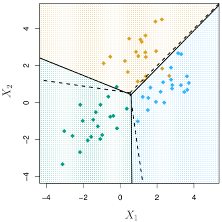
LDA - $p > 1$

- Assume observations in the kth drawn from multi-variate normal distribution with the same covariance matrix for all k: $N(\mu_k, \Sigma)$
- For new date point x, predict the k for which $p_k(x) = \Pr(Y = k|X = x)$ is largest
- Equivalent to finding k for which $\delta_k(x)$ is largest

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

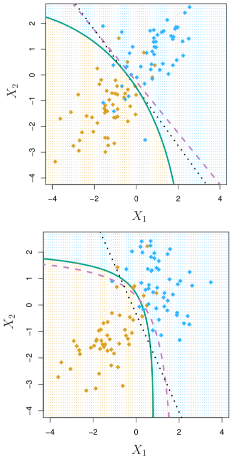
- Approximate μ_i's π_i's and Σ to find boundary where the returned k switches





Quadratic Discriminant Analysis (QDA)

- Same idea as LDA, but don't assume same covariance matrix
- Assume observations in kth colass drawn from normal distribution $N(\mu_k, \Sigma_k)$
- Make new predictions based on $\delta_k(x) = \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$
- This setup means decision boundaries are quadratic



Example

Assume the probability that a person defaulted on credit card payment is 10%. Based on the given use LDA to predict the default status of a person with a credit card balance of 1800.

Balance	**Prediction**
0	*No*
500	*No*
1000	*Yes*
1500	*No*
2000	*Yes*
2500	*Yes*

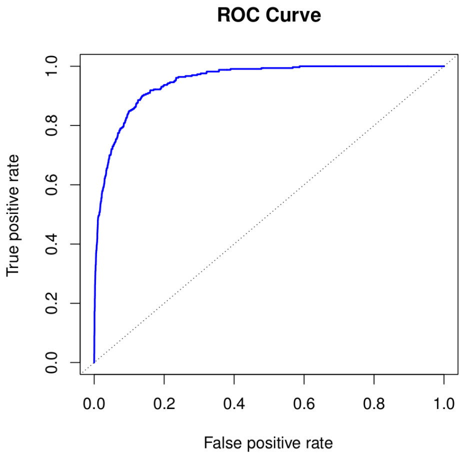
Confusion matrix and types of Errors

		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

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

ROC curve


ROC Curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Multivariate normal distribution

Gaussian (normal) distributions

- $Z \sim N(0,1)$ means Z follows a standard Gaussian distribution, i.e., has probability density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- If $Z_1, Z_2, ..., Z_d$ are iid $N(0,1)$ random variables, then say $\mathbf{Z} := (Z_1, Z_2, ..., Z_d)$ follows a standard multivariate Gaussian distribution on $\mathbb{R}^d$, i.e., $\mathbf{Z} \sim N(0, I_d)$.

- Other Gaussian distributions on $\mathbb{R}^d$ arise by applying (invertible) linear maps and translations to $\mathbf{Z}$:

$$\mathbf{z} \mapsto \mathbf{Az} \mapsto \mathbf{Az} + \boldsymbol{\mu}.$$

  - $\mathbf{X} := \mathbf{AZ} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \mathbf{AA}^\mathsf{T})$ has

$$\mathbb{E}(\mathbf{X}) = \mu \quad \text{and } \operatorname{cov}(\mathbf{X}) = \mathbf{AA}^\mathsf{T}$$

  - the (i,j)th entry of $\operatorname{cov}(\mathbf{X})$ represents the correlation between X_i and X_j.

Appendix: Review of Multivariate normal distribution

- $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has the probability density function

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from data: maximum likelihood estimators
  - $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$
  - $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$
  - $\hat{\boldsymbol{\mu}}$ is unbias and $\hat{\boldsymbol{\Sigma}}$ is slightly biased