

Classification (Cont.)

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

Recap

- ▶ We observe (y_i, X_i) $i = 1, \dots, n$
- ▶ Estimate Probabilities
 - ▶ Logit

$$p_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \quad (1)$$

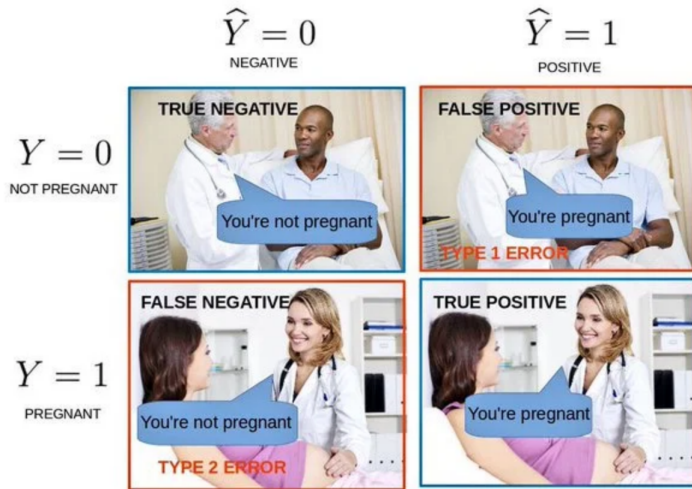
- ▶ get β
- ▶ Prediction
 - ▶ Logit, with the $\hat{\beta}$

$$\hat{p}_i = \frac{e^{X_i \hat{\beta}}}{1 + e^{X_i \hat{\beta}}} \quad (2)$$

- ▶ Classification

$$\hat{Y}_i = 1[\hat{p}_i > c] \quad (3)$$

Performance



Performance

		\hat{y}_i	
		0	1
y_i	0	TN	FP
	1	FN	TP

- ▶ We have two types of error associated with this that we can use as a measure of performance

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{Negatives}}$$

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{Positives}} \quad (4)$$

- ▶ Another names they receive:
 - ▶ False positive rate: Type I error, 1-Specificity
 - ▶ True positive rate: 1- Type II error, power, sensitivity.

Agenda

① Recap

② Other Models for Classification

- Regularization for Logit
- KNN
- Discriminant Analysis
- Naive Bayes

③ Extra: Kappa statistic

Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

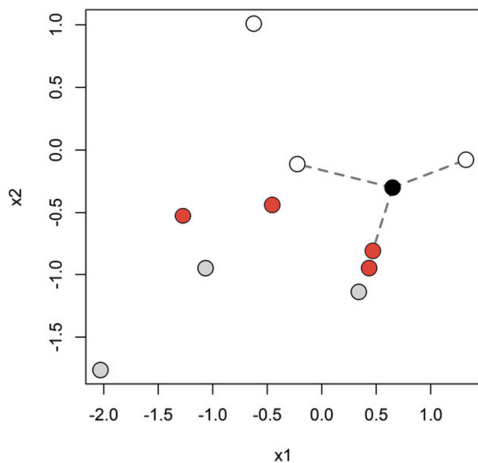
Regularization for Logit

Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

K-Nearest Neighbors

- K nearest neighbor (K-NN) algorithm predicts class \hat{y} for x by asking *What is the most common class for observations around x ?*



K-Nearest Neighbors

- ▶ K nearest neighbor (K-NN) algorithm predicts class \hat{y} for x by asking *What is the most common class for observations around x ?*
- ▶ Algorithm: given an input vector x_f where you would like to predict the class label
 - ▶ Find the K nearest neighbors in the dataset of labeled observations, $\{x_i, y_i\}_{i=1}^n$, the most common distance is the Euclidean distance:

$$d(x_i, x_f) = \sqrt{\sum_{j=1}^p (x_{ij} - x_{fj})^2} \quad (5)$$

- ▶ This yields a set of the K nearest observations with labels:

$$[x_{i1}, y_{i1}], \dots, [x_{iK}, y_{iK}] \quad (6)$$

- ▶ The predicted class of x_f is the most common class in this set

$$\hat{y}_f = \text{mode}\{y_{i1}, \dots, y_{iK}\} \quad (7)$$

K-Nearest Neighbors

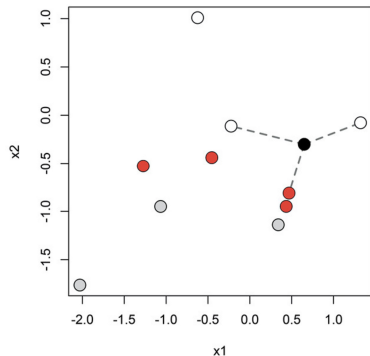
- There are some major problems with practical implications
 - Knn predictions are unstable as a function of K

$$K = 1 \implies \hat{p}(white) = 0$$

$$K = 2 \implies \hat{p}(white) = 1/2$$

$$K = 3 \implies \hat{p}(white) = 2/3$$

$$K = 4 \implies \hat{p}(white) = 1/2$$



Source: Taddy (2019)

K-Nearest Neighbors

- ▶ There are some major problems with practical implications
 - ▶ Knn predictions are unstable as a function of K
 - ▶ This instability of prediction makes it hard to choose the optimal K and cross validation doesn't work well for KNN
 - ▶ Since prediction for each new x requires a computationally intensive counting, KNN is too expensive to be useful in most big data settings.
 - ▶ KNN is a good idea, but too crude to be useful in practice

Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

Linear Discriminant Analysis

Reverend Bayes to the rescue: Bayes Theorem

$$Pr(Y = 1|X) = \frac{f(X|Y = 1)\pi(Y = 1)}{m(X)} \quad (8)$$

with $m(X)$ is the marginal distribution of X , i.e.

$$m(X) = \int_y f(X|Y = y)\pi(Y = y)dy \quad (9)$$

Linear Discriminant Analysis

Reverend Bayes to the rescue: Bayes Theorem

Recall that there are two states of nature $y \rightarrow i \in \{0, 1\}$

$$m(X) = f(X|Y = 1)\pi(Y = 1) + f(X|Y = 0)\pi(Y = 0) \quad (10)$$

$$m(X) = f(X|Y = 1)\pi(Y = 1) + f(X|Y = 0)(1 - \pi(Y = 1)) \quad (11)$$

Linear Discriminant Analysis

- ▶ This is basically an empirical Bayes approach
- ▶ We need to estimate $f(X|Y = 1), f(X|Y = 0)$ and $\pi(Y = 1)$
 - ▶ Let's start by estimating $\pi(Y = 1)$. We've done this before

$$\pi(Y = 1) = \frac{\sum_{i=1}^n 1[Y_i = 1]}{N} \quad (12)$$

Linear Discriminant Analysis

- ▶ Next $f(X|Y = j)$ with $j = 0, 1$.
 - ▶ if we assume one predictor and $X|Y \sim N(\mu_j, \sigma_j)$, the problem boils down to estimating μ_j, σ_j
 - ▶ LDA makes it simpler, assumes $\sigma_j = \sigma \forall j$
 - ▶ then partition the sample in two $Y = 0$ and $Y = 1$, estimate the moments and get $\hat{f}(X|Y = j)$
- ▶ Plug everything into the Bayes Theorem and you're done

Linear Discriminant Analysis

Extensions

- ▶ If we have k predictors?
- ▶ then $X|Y \sim NM(\mu, \Sigma)$

$$f(X|Y = j) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_j)' \Sigma_j (x - \mu_j)\right) \quad (13)$$

- ▶ μ_j is the vector of the sample means in each partition $j = 0, 1$
- ▶ Σ_j is the matrix of variance and covariances of each partition $j = 0, 1$

Linear Discriminant Analysis

- ▶ Why is it called linear?
- ▶ Note

$$p > \frac{1}{2} \iff \ln\left(\frac{p}{(1-p)}\right) \quad (14)$$

- ▶ Logit with one predictor

$$\beta_1 + \beta_2 X \quad (15)$$

- ▶ Classification: in the probability of space
- ▶ Discrimination: in the space of X
- ▶ $\beta_1 + \beta_2 X$ is the discrimination function for logit (it is lineal)

Linear Discriminant Analysis

- ▶ LDA?
- ▶ One predictor with $\sigma_0 = \sigma_1$ (equal variance)

$$Pr(Y = 1|X) = \frac{f(X|Y = 1)\pi(Y = 1)}{f(X|Y = 1)\pi(Y = 1) + f(X|Y = 0)(1 - \pi(Y = 1))} \quad (16)$$

- ▶ Then under the equal variance assumption

$$\frac{Pr(Y = 1|X)}{1 - Pr(Y = 1|X)} = \frac{f(X|Y = 1)\pi(Y = 1)}{f(X|Y = 0)(1 - \pi(Y = 1))} \quad (17)$$

$$= \frac{\pi(Y = 1)\exp(-(x - \mu_1)^2)}{(1 - \pi(Y = 1))\exp(-(x - \mu_0)^2)} \quad (18)$$

Linear Discriminant Analysis

- ▶ Taking logs

$$\log \left(\frac{Pr(Y = 1|X)}{1 - Pr(Y = 1|X)} \right) = \log \left(\frac{\pi(Y = 1)}{(1 - \pi(Y = 1))} + (x - \mu_1)^2 - (x - \mu_0)^2 \right) \quad (19)$$

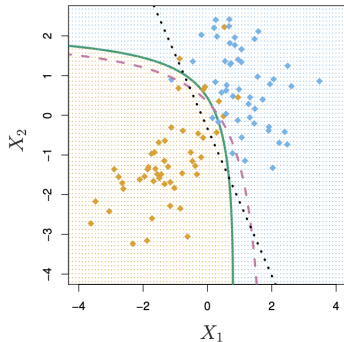
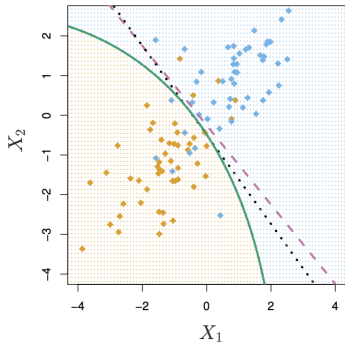
$$= \log \left(\frac{\pi(Y = 1)}{(1 - \pi(Y = 1))} + \mu_1^2 - \mu_0^2 - 2(\mu_1 - \mu_0)x \right) \quad (20)$$

$$= \gamma_1 + \gamma_2 X \quad (21)$$

- ▶ under the assumption of equal variance the discrimination function is lineal
- ▶ Note: logit estimates γ_1 and γ_2

Quadratic Discriminant Analysis

- QDA assumes different variances for the components



Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

Naive Bayes

$$Pr(Y = 1|X) = \frac{f(X|Y = 1)\pi(Y = 1)}{f(X|Y = 1)\pi(Y = 1) + f(X|Y = 0)(1 - \pi(Y = 1))} \quad (22)$$

- ▶ $\pi(Y = 1)$
- ▶ $f(X|Y = 1)$

Naive Bayes

- NB assumes independence

$$f(X|Y = 1) = f(x_1|Y = 1) \times \dots \times f(x_k|Y = 1) \quad (23)$$

Example: Default



photo from <https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/>

Agenda

- 1 Recap
- 2 Other Models for Classification
 - Regularization for Logit
 - KNN
 - Discriminant Analysis
 - Naive Bayes
- 3 Extra: Kappa statistic

Kappa statistic

- ▶ Also known as Cohen's Kappa
- ▶ It was originally designed to assess the agreement between two raters (Cohen 1960).
- ▶ Kappa takes into account the accuracy that would be generated simply by chance.

$$Kappa = \frac{O - E}{1 - E} \quad (24)$$

- ▶ Take on values between -1 and 1;
- ▶ 0 means no agreement between the observed and predicted classes,
- ▶ 1 indicates perfect concordance of the model prediction and the observed classes.
- ▶ Negative values indicate that the prediction is in the opposite direction of the truth