Classification (Cont.)

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

Recap

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

$$p_i = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \tag{1}$$

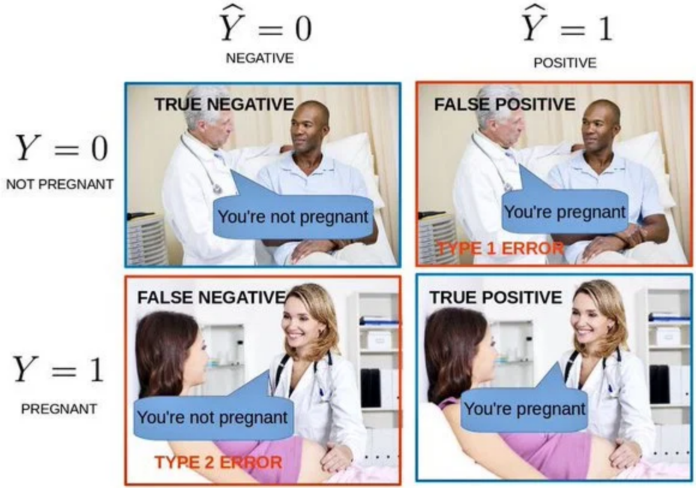
  - get β
- Prediction
  - Logit, with the $\hat{\beta}$

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

- Classification

$$\hat{Y}_i = 1[\hat{p}_i > c] \tag{3}$$

Performance



Performance

$$
\begin{array}{cccc}
 & & \hat{y}_i & \\
 & & 0 & 1 \\
\multirow{2}{*}{y_i} & 0 & TN & FP \\
 & 1 & FN & TP
\end{array}
$$

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

$$
\begin{aligned}
False\ Positive\ Rate &= \frac{False\ \ Positives}{Negatives} \\
True\ Positive\ Rate &= \frac{True\ \ Positives}{Positives}
\end{aligned} \tag{4}
$$

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

Agenda

Agenda

Regularization for Logit

$$\min_{\beta_0,\ldots,\beta_k} \frac{1}{N}\sum_{i=1}^{N} l(y_i, \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k) + \lambda\left(\alpha\sum_{j=1}^{k}|\beta_j| + (1-\alpha)\ \sum_{j=1}^{k}(\beta_j)^2/2\right) \tag{5}$$

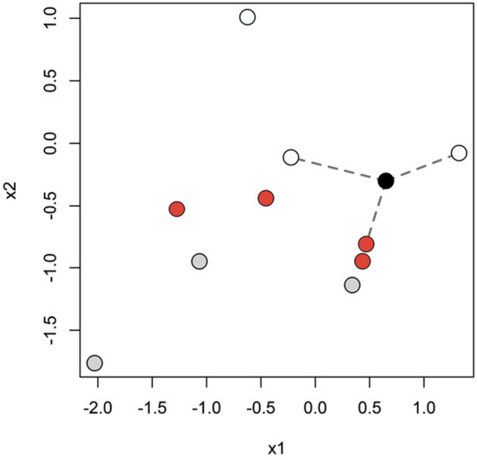
- Si $\alpha = 1$ Lasso
- Si $\alpha = 0$ Ridge

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?*



Source: Taddy (2019)

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:

$$d(x_i, x_f) = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{fj})^2} \tag{6}$$

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

$$[x_{i1}, y_{i1}], \ldots, [x_{iK}, y_{iK}] \tag{7}$$

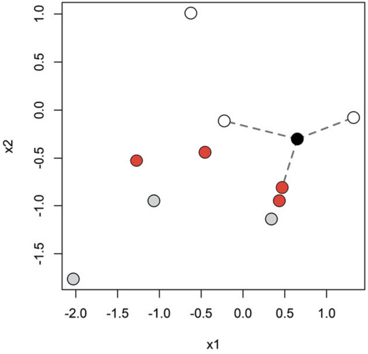
  - The predicted class of x_f is the most common class in this set

$$\hat{y}_f = mode\{y_{i1}, \ldots, y_{iK}\} \tag{8}$$

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

Linear Discriminant Analysis

Reverend Bayes to the rescue: Bayes Theorem

$$Pr(Y=1|X) = \frac{f(X|Y=1)\pi(Y=1)}{m(X)} \tag{9}$$

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

$$m(X) = \int_y f(X|Y=y)\pi(Y=y)dy \tag{10}$$

Linear Discriminant Analysis

Reverend Bayes to the rescue: Bayes Theorem

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

$$m(X) = f(X|Y=1)\pi(Y=1) + f(X|Y=0)\pi(Y=0) \tag{11}$$

$$m(X) = f(X|Y=1)\pi(Y=1) + f(X|Y=0)(1-\pi(Y=1)) \tag{12}$$

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

$$\pi(Y=1) = \frac{\sum_{i=1}^{n} 1[Y_i = 1]}{N} \tag{13}$$

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 $X|Y \sim NM(\mu, \Sigma)$

$$f(X|Y=j) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu_j)'\Sigma_j(x-\mu_j) \tag{14}$$

- 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

$$p > \frac{1}{2} \iff ln(\frac{p}{(1-p)}) \tag{15}$$

- Logit with one predictor

$$\beta_1 + \beta_2 X \tag{16}$$

- 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)

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

- Then under the equal variance assumption

$$\frac{Pr(Y=1|X)}{1-Pr(Y=1|X)} = \frac{f(X|Y=1)\pi(Y=1)}{f(X|Y=0)(1-\pi(Y=1))} \tag{18}$$

$$= \frac{\pi(Y=1)exp((x-\mu_1)^2)}{(1-\pi(Y=1))exp((x-\mu_0)^2)} \tag{19}$$

Linear Discriminant Analysis

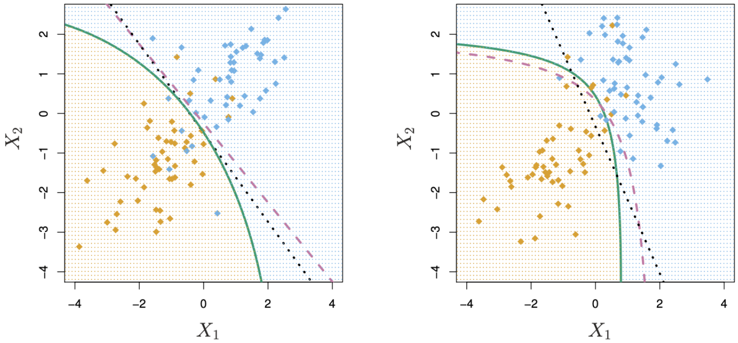
- Taking logs

$$
\begin{aligned}
log\left(\frac{Pr(Y=1|X)}{1-Pr(Y=1|X}\right) &= log(\frac{\pi(Y=1)}{(1-\pi(Y=1))} + (x-\mu_1)^2 - (x-\mu_0)^2 && (20)\\
&= log(\frac{\pi(Y=1)}{(1-\pi(Y=1))} + \mu_1^2 - \mu_0^2 - 2(\mu_1-\mu_0)x && (21)\\
&= \gamma_1 + \gamma_2 X && (22)
\end{aligned}
$$

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

Quadratic Discriminant Analysis

- QDA assumes diferent variances for the components



Agenda

Naive Bayes

$$Pr(Y = 1|X) = \frac{f(X|Y = 1)\pi(Y = 1)}{f(X|Y = 1)\pi(Y = 1) + f(X|Y = 0)(1 - \pi(Y = 1))} \tag{23}$$

- $\pi(Y = 1)$
- $f(X|Y = 1)$

Naive Bayes

- NB assumes independence

$$f(X|Y=1) = f(x_1|Y=1) \times \cdots \times f(x_k|Y=1) \tag{24}$$

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
- I 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.

$$Kappa = \frac{O - E}{1 - E} \tag{25}$$

- 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