Supervised Learning - linear classification

Mark Asch - IMU/VLP/CSU

2023

Program

1. Data Analysis

   (a) Introduction: the 4 identifiers of "big data" and "data science"
   (b) Supervised learning methods: regression—advanced, k-NN, linear classification methods, SVM, NN, decision trees.
   (c) Unsupervised learning methods: k-means, principal component analysis, clustering.

Methods

1. Logistic Regression (classification)
2. Bayes Classifier
3. LDA (Linear Discriminant Analysis)

Introduction

- Classification problems are very widespread.

  $\Rightarrow$ life is full of binary choices...

Logistic Regression

- In spite of its name, this is actually a classification method.

- The reasons for its popularity are:

  ✔ easy to implement and deploy
  ✔ easy to interpret
  ✔ very efficient training
  ✔ very fast classification of new data
  ✔ can provide information on the importance of features

- There are, however, three limitations.

  ✘ a linear hypothesis where the odds (see below) are linearly dependent on the predictors;
  ✘ the frontier between 2 classes is linear;
  ✘ only valid for binary classification, i.e. cases where there are only two classes.

Logistic Regression II

- even though it is used for classification... we suppose that:

  $\Rightarrow$ we have a binary response, yes or no, malignant or benign, sick or healthy, alive or dead. . . taking the value 0 or 1.

  $\Rightarrow$ and that we want to model the response (conditional probability)

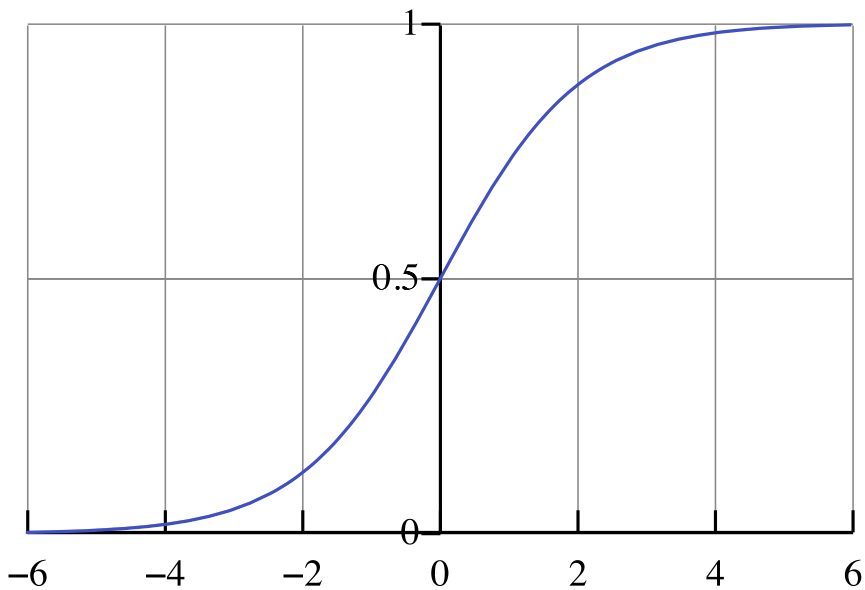
$$p(X) \doteq P(Y = 1 \mid X)$$

- for this we use the logistic function

The logistic function

Definition. The logistic function (sigmoid) is a mapping from $\mathbb{R}$ into $[0, 1]$ defined by

$$p(X) = \frac{e^X}{1 + e^X} = \frac{1}{1 + e^{-X}}$$



- Suppose now that we have a linear model for X of the form

$$\beta_0 + \beta_1 X,$$

then the logistic function becomes

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

and so

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

which is the odds ratio

- Taking the logarithm, we get the logit function

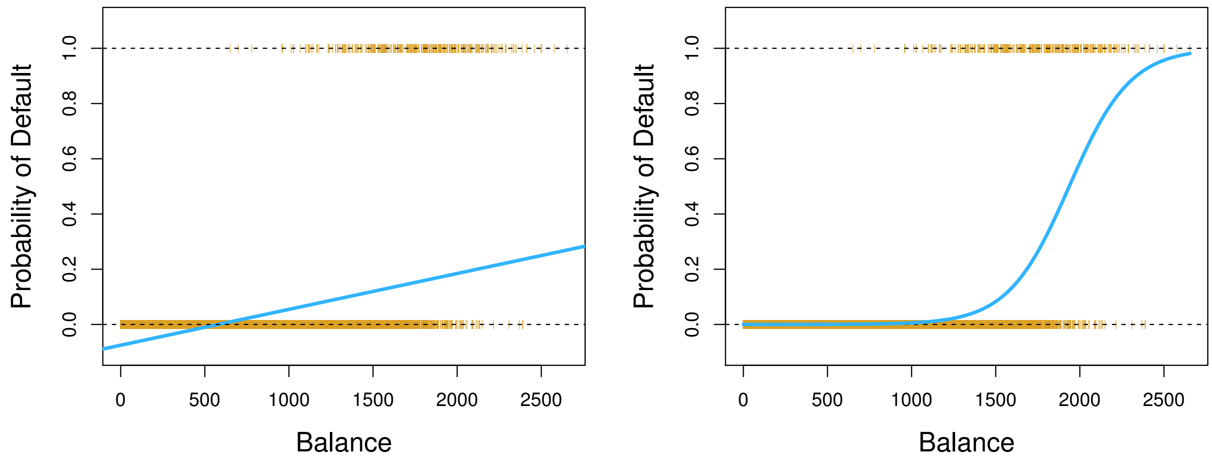
$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X$$

⇒ An increase of one unit in X produces an increase of $\boldsymbol{\beta_1}$ units in $p(X)$.

⇒ The coefficients $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$ are estimated by a maximum likelihood method

- Prediction: for a new, unseen value of X, we have the estimation

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X)}}$$



- The logistic model can be extended to several predictors $X_1, \ldots, X_p$ but not to more than 2 classes.

  $\Rightarrow$ for this we will use the LDA (or a nonlinear method such as SVM, etc.)

Linear Discriminant Analysis (LDA)

- Linear discriminant analysis extends logistic regression to the case where we have more than two classes.

- We saw that LR models the conditional probability,

$$P(Y = k \mid X = x)$$

  and that logistic regression models this probability directly

  $\Rightarrow$ using the logistic/sigmoid function, and
  $\Rightarrow$ for the case of two response classes (binary)

- For several classes, we must use Bayes' Law to compute the desired conditionals.

  $\Rightarrow$ we need to model the distribution of the predictors separately for each class, and then

⇒ use Bayes' Law to estimate the desired conditionals $P(Y = k \mid X = x)$, as follows

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)},$$

where

→ π_k is the prior probability of class $k = 1, \ldots, K$

→ $f_k(x) = P(X = x \mid Y = k)$ is the likelihood

→ $p_k(x) = P(Y = k \mid X = x)$ is the posterior probability that the observation is of class k given the value of the predictor $X = x$

- In LDA we suppose

⇒ $f_k(x) \sim \mathcal{N}(\mu_k, \sigma_k)$ is Gaussian

⇒ the variances σ_k are equal

- This gives the theoretical class frontier, known as the Bayes classifier,

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

also called the discriminant function, linear x,

- Then simply affect each observation to the class k for which this value is maximal.

- Finally, the LDA classifier is the approximation

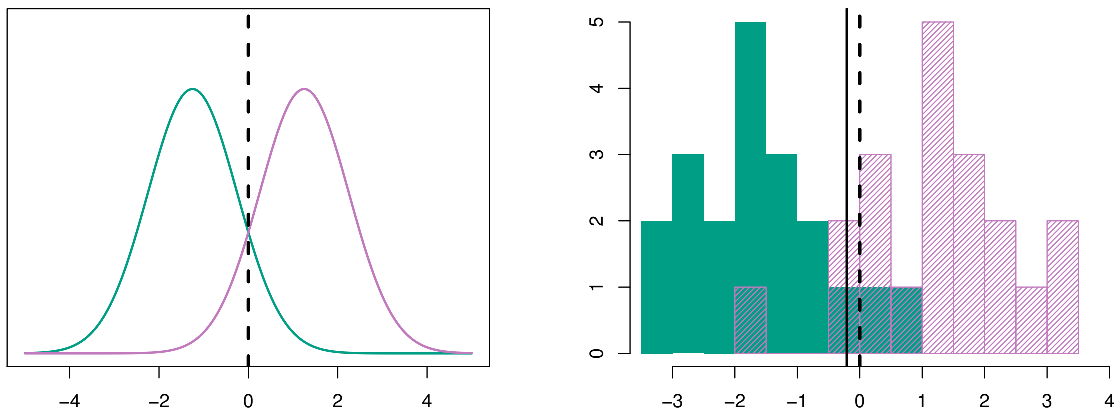
$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k$$

where

$\Rightarrow$ $\hat{\pi}_k = n_k/n$
$\Rightarrow$ $\hat{\mu_k} = (1/n_k)\sum_{i:y_i=k} x_i$
$\Rightarrow$ $\hat{\sigma}^2 = 1/(n-K)\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i-\mu_k)^2$

LDA - example

- An example for classifying 2 Gaussian distributions:



- Left : 2 normal distributions, Bayes decision boundary (dashed line)

- Right : 20 observations drawn from each class, LDA decision boundary (solid line)

$\Rightarrow$ $n_1 = n_2$, so $\hat{\pi}_1 = \hat{\pi}_2$ and the decision boundary is at $(\hat{\mu}_1 + \hat{\mu}_2)/2$.

Naive Bayes Classifier (NB)

- A family of supervised classifiers based on
  - ⇒ Bayes' Theorem
  - ⇒ the naive hypothesis of pairwise conditional independence of the features, knowing the value of the class variable

- Recall Bayes formula

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

- Naive hypothesis

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

and thus Bayes becomes

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

- Classification rule is then

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

since

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

- We use the MAP approximation to estimate $P(y)$ and $P(x_i \mid y)$

- The classifiers differ in the form of the distribution of $P(x_i \mid y)$

  $\Rightarrow$ Gaussian - for continuous values
  $\Rightarrow$ Bernoulli - for binary outcomes
  $\Rightarrow$ Multinomial - for the number or the frequency of an outcome

NB: pros and cons

✔ Robustness in presence of noisy or missing data.

✔ Resistance to overfitting.

✔ Efficiency for small samples.

✘ Bad estimation of probabilities...

How to choose a model?

- Once the tuning parameters determined, we still must choose between several models

  $\Rightarrow$ the choice will largely depend on the data characteristics and the type of questions we ask

- But, predicting which model will be the most pertinent, is in general quite difficult...

- A recommended scheme for finalizing the choice is as follows:

1. Begin with a few models that are the least interpretable and the most flexible. These models have a strong chance of producing more precision.

   (a) SVM
   (b) Trees with boosting.
   (c) Random Forests (RF)

2. Study simpler models, that are less opaque.
    (a) Linear models.
    (b) (GAM/GLM)
    (c) Naive Bayes.
    (d) k-NN.
    (e) Logistic Regression.
    (f) Regression Splines (MARS).

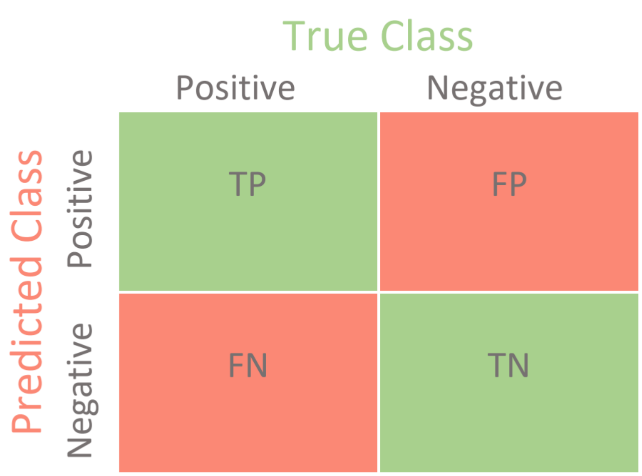
3. Use, if possible, the simplest model that approximates reasonably well the performance of the more complex models.

Confusion Tables and ROC Curves

- How can one evaluate and display the precision of a classification method?

  $\Rightarrow$ Confusion Tables
  $\Rightarrow$ ROC curves

- Please see the Advanced Course lecture for further details.



Definition 1. A *confusion table/matrix* , C, for a classi-

fication with n classes is an $(n \times n)$ matrix with entries

$$C_{ij} = \text{the number of observations in class } i$$
$$\text{that are predicted in class } j.$$

Then,

- C_{ii}, $i = 1, \ldots, n$, are the good classifications,
- C_{ij} with $i \neq j$ are the bad classifications.
- The global precisionis defined as

$$\frac{TP + TN}{TP + TN + FP + FN} = \frac{\sum_{i=1}^{n} C_{ii}}{\sum_i \sum_j C_{ij}}$$

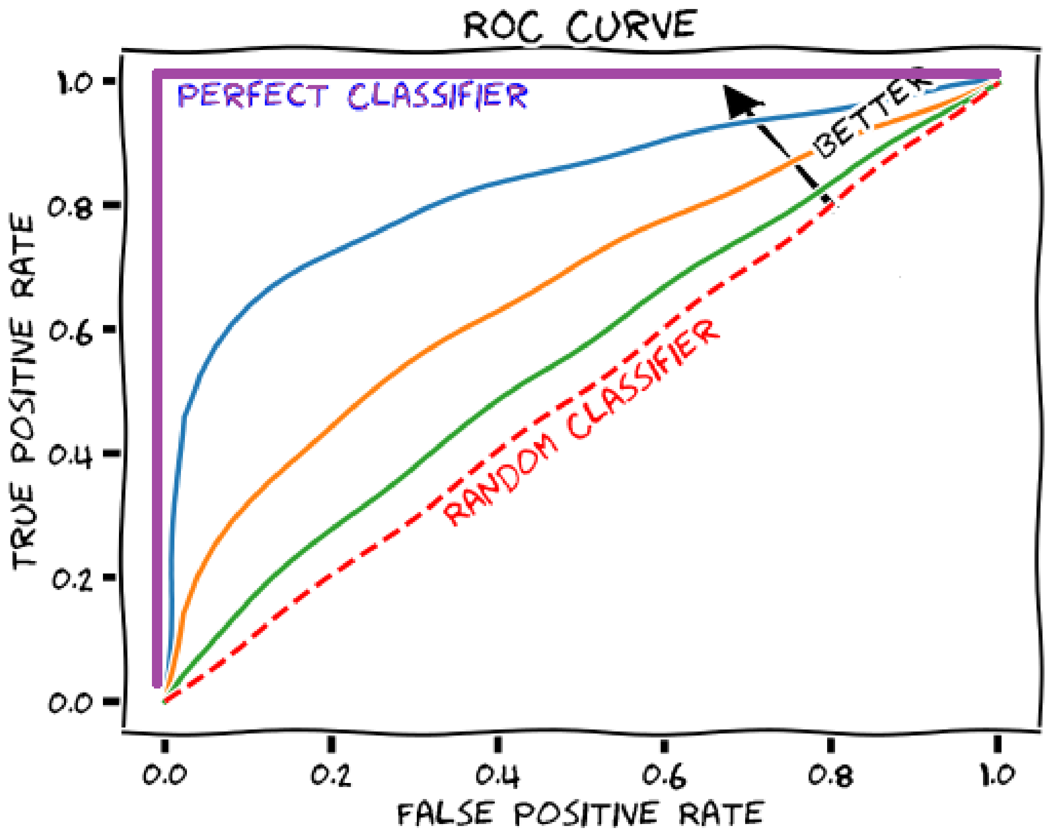
- ROC (Receiver Operating Curve)

  $\Rightarrow$ a parametric curve with classification thresholds
  $\Rightarrow$ better because robust to unbalanced datasets, not having the same number of samples in each class

⇒ the two axes are the true positive rate (TPR) and the false positive rate (FPR), where

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN}$$

⇒ often, we compute the AUC which is the area under the ROC, and is scale invariant


ROC CURVE
PERFECT CLASSIFIER
BETTER
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
0.0
0.2
0.4
0.6
0.8
1.0

Examples

1. Logistic Regression for prediction of hurricane class - `reg-logistic.html`

2. LDA Classification - `lda_caret_iris.html`

3. Naive Bayes Classification - `NB_caret.html`