Classification

- Used to handle categorical (qualitative) data:
  - Eg: ① Spam or Not Spam
    ② Male or Female
    ③ Eye color ∈ {Blue, Green, Brown}

- Classification majorly revolves around probabilities (prob. that X belong to each category in C)

- The class 'k' is assigned to x, for which probability is maximum.

example:

Case: Credit Card Default Data

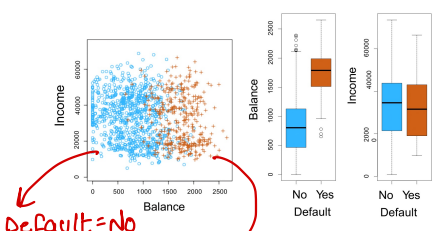
- To predict customers that are likely to default
- Possible *X* variables are:
  - Annual Income
  - Monthly credit card balance
- The *Y* variable (Default) is categorical: Yes or No
- How do we check the relationship between *Y* and *X*?

p = 2

k = 2

5

Example: Credit Card Defualt



6

★ Classification using linear regression:

Can we use Linear Regression?

Suppose for the Default classification task that we code

$$Y = \begin{cases} 0 & \text{if No} \\ 1 & \text{if Yes.} \end{cases}$$

Can we simply perform a linear regression of *Y* on *X* and classify as Yes if $\hat{Y} > 0.5$?

7

In case of binary classification, we can use LDA.

- LR cannot be used if it is not a binary classification.

Linear Regression continued

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure} \end{cases}$$

This coding suggests an ordering, and in fact implies that the difference between stroke and drug overdose is the same as between drug overdose and epileptic seizure.

11

Multi-Class and Multi-Label Problems

Multiclass classification means a classification task with more than two classes; e.g., classify a set of images of animals which may be horses, birds, or fish.

eg:

Multiclass classification makes the assumption that each sample is assigned to one and only one label: an animal can be either a horse or a bird but not both at the same time.

13

Multi-Class and Multi-Label Problems

Multilabel classification assigns to each sample a set of target labels. This can be thought as predicting properties of a data-point that are not mutually exclusive, such as topics that are relevant for a document.

eg.:

A text might be about any of religion, politics, finance or education at the same time or none of these.

14

Binary Classification

K=2

A binary classification task assigns only one of the two possible classes to each observation.

Because multi-class and multi-label classification tasks can be performed using binary classification techniques, many times we focus on binary classification.

15

Logistic Regression ensures that p(x) lies betⁿ 0 and 1. (log funcⁿ)

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using balance to predict default. Logistic regression uses the form

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

($e \approx 2.71828$ is a mathematical constant [Euler's number.])

It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

16

Credit card example

Y = 1 ⇒ Default = Yes

Y = 0 ⇒ Default = No

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

A bit of rearrangement gives

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

This monotone transformation is called the *log odds* or *logit* transformation of $p(X)$.

18

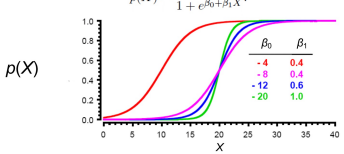
Sigmoid Function

Parameters control shape and location of sigmoid curve

β_0 controls location of midpoint

β_1 controls slope of rise

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



When x=- β_0/β_1 , $\beta_0+\beta_1 x=0$; thus $p(X) = 0.5$

17

Definition of Odds

- The probability of an event divided by the probability of its complement is called its odds.
- Example: The probability of winning in a casino is 1%. What is the odds of winning in that casino?

O(W)=Pr(W)/(1-Pr(W))
=0.01/0.99=1/99

19

Logit Model

Therefore, the logit model is trying to predict the log of odds of a model as a linear combination of the predictor(s):

$\log(O(Y=1|X=x))=\beta_0+\beta_1 X$

20

Class as A Bernoulli Random Variable

One can see class in a binary classification problem as a Bernoulli random variable that can take two values 0 and 1:

$$\Pr(Y=1|X=x) = p(x)$$
$$\Pr(Y=0|X=x) = 1-p(x)$$

This can be rewritten as:

$$p_{Y|X}(y|x) = \Pr(Y=y|X=x) = [p(x)]^{y}[1-p(x)]^{1-y}, y=0,1$$

22

Independent Sample of Bernoulli Variables

Assume that we have an *independent* sample whose classes are $Y^{(1)} = y_1$, $Y^{(2)} = y_2, \ldots, Y^{(N)} = y_N$

y_i is 0 or 1.

The *joint probability mass* function of this independent sample is:

$p_{Y|X}(y_1, y_2, \ldots, y_N|\text{Data})$
$= \Pr(Y^{(1)} = y_1, Y^{(2)} = y_2, \ldots, Y^{(N)} = y_N|\text{Data})$
$= \Pr(Y^{(1)} = y_1|\text{Data})\ \Pr(Y^{(2)} = y_2|\text{Data}) \ldots \Pr(Y^{(N)} = y_N|\text{Data})$
Data: $X^{(1)} = x_1, \ldots, X^{(N)} = x_N$

23

✡ Likelyhood Function :

Independent Sample of Bernoulli Variables

The joint probability mass function is a function of β_0, β_1 and is called the likelihood function, given the data samples.

Maximum likelyhood can be calculated using : ① Gradient method ② Expectation maximization

25

Maximum Likelihood

We use maximum likelihood to estimate the parameters β_0, β_1.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1-p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

30

✡ Hypothesis Testing :

$H_0 = 0 \Rightarrow$ NULL
$H_0 \neq 0 \Rightarrow H_A \Rightarrow$ Reject Null

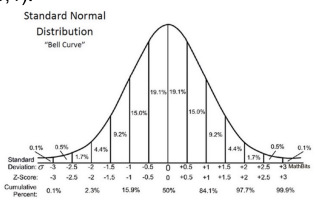
- Instead of t-test, we use Z-test
- Both are same

$$z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

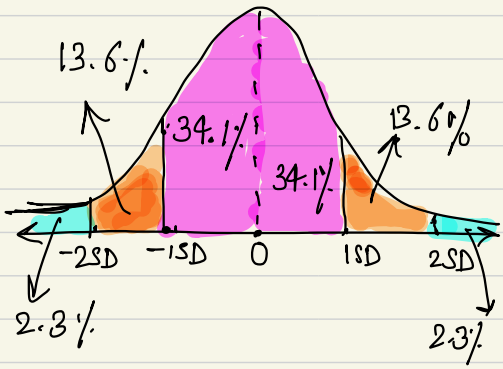
Are the coefficients significant?

The quantity $z = \frac{\hat{\beta}_i - \beta}{\mathrm{SE}(\hat{\beta}_i)}$

follows a *standard normal distribution* $N(0,1)$.



35



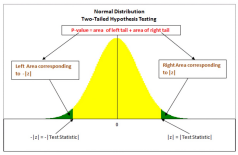
- For $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$:

reject NULL hypothesis.

if p is small, reject NULL.

Are the coefficients significant?

The p-value is the probability of observing something whose magnitude is bigger than |z|:



39

Multiple Logistic Regression:

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

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

Class Imbalance:

Class Imbalance

- Intuitively, a dataset is imbalanced when members of certain class(es) are rare.
- The lack of observations of certain classes does not always imply their irrelevance.
- For example, in medical studies of rare diseases, the small number of infected patients (cases) conveys the most valuable information for diagnosis and treatments.

Types of Imbalance

•Formally, an imbalanced dataset exhibits one or more of the following properties:

•**Marginal Imbalance**. A dataset is marginally imbalanced if one class is rare compared to the other class. In other words, $\Pr(Y=1) \approx 0$.

Types of Imbalance

•Formally, an imbalanced dataset exhibits one or more of the following properties:

- ***Conditional Imbalance***. A dataset is conditionally imbalanced when it is easy to predict the correct labels in most cases. For example, if $X \in \{0,1\}$, the dataset is conditionally imbalanced if $\Pr(Y=1|X=0) \approx 0$ and $\Pr(Y=1|X=1) \approx 1$.

→ Upsampling
→ Downsampling
→ SMOTE

} To read

✡ Multiclass Logistic Regression :

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell}+\beta_{1\ell}X_1+\ldots+\beta_{p\ell}X_p}}$$

Aka multinomial Logistic Regression.

67

✡ Bayesian Discriminant Analysis :

- In this approach, we model distribution of X into all K classes and flip the probabilities using Bayes Theorem.

+ Bayes Theorem :

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

OR

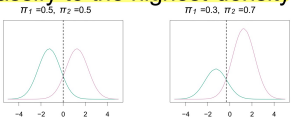
$$Pr(Y=K|X=x) = \frac{\pi_K f_K(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

72

where $f_K(x) = Pr(X=x|Y=K)$

π_K = Marginal prob. OR Prior Probability

Classify to the highest density



- We classify a new point according to which density is highest.
- When the priors are different, we take them into account as well, and compare $\pi_k f_k(x)$. On the right, we favor the pink class — the decision boundary has shifted to the left.

74

Softmax Coding

✡ More on Multinomial Logistic Regression :

4.3.5 Multinomial Logistic Regression

We sometimes wish to classify a response variable that has more than two classes. For example, in Section 4.2 we had three categories of medical condition in the emergency room: `stroke`, `drug overdose`, `epileptic seizure`. However, the logistic regression approach that we have seen in this section only allows for $K = 2$ classes for the response variable.

It turns out that it is possible to extend the two-class logistic regression approach to the setting of $K > 2$ classes. This extension is sometimes known as *multinomial logistic regression*. To do this, we first select a single class to serve as the *baseline*; without loss of generality, we select the Kth class for this role. Then we replace the model (4.7) with the model

multinomial logistic regression

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \quad (4.10)$$

for $k = 1,\ldots,K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \quad (4.11)$$

It is not hard to show that for $k = 1,\ldots,K-1$,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p. \quad (4.12)$$

Notice that (4.12) is quite similar to (4.6). Equation 4.12 indicates that once again, the log odds between any pair of classes is linear in the features.

It turns out that in (4.10)–(4.12), the decision to treat the Kth class as the baseline is unimportant. For example, when classifying emergency room visits into `stroke`, `drug overdose`, and `epileptic seizure`, suppose that we fit two multinomial logistic regression models: one treating `stroke` as the baseline, another treating `drug overdose` as the baseline. The coefficient estimates will differ between the two fitted models due to the differing choice of baseline, but the fitted values (predictions), the log odds between any pair of classes, and the other key model outputs will remain the same.

Nonetheless, interpretation of the coefficients in a multinomial logistic regression model must be done with care, since it is tied to the choice of baseline. For example, if we set `epileptic seizure` to be the baseline, then we can interpret $\beta_{\texttt{stroke}0}$ as the log odds of `stroke` versus `epileptic seizure`, given that $x_1 = \ldots = x_p = 0$. Furthermore, a one-unit increase in X_j is associated with a $\beta_{\texttt{stroke}j}$ increase in the log odds of `stroke` over `epileptic seizure`. Stated another way, if X_j increases by one unit, then

Imp.

$$\frac{\Pr(Y = \texttt{stroke}|X = x)}{\Pr(Y = \texttt{epileptic seizure}|X = x)}$$

increases by $e^{\beta_{\texttt{stroke}j}}$.

We now briefly present an alternative coding for multinomial logistic regression, known as the *softmax* coding. The softmax coding is equivalent to the coding just described in the sense that the fitted values, log odds between any pair of classes, and other key model outputs will remain the same, regardless of coding. But the softmax coding is used extensively in some areas of the machine learning literature (and will appear again in Chapter 10), so it is worth being aware of it. In the softmax coding, rather than selecting a baseline class, we treat all K classes symmetrically, and assume that for $k = 1, \ldots, K$,

softmax

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{4.13}$$

Thus, rather than estimating coefficients for $K - 1$ classes, we actually estimate coefficients for all K classes. It is not hard to see that as a result of (4.13), the log odds ratio between the kth and k'th classes equals

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p. \tag{4.14}$$

Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data.

LDA

Linear Discriminant Analysis ($p = 1$)

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

where, σ_k^2 = variance
σ_k = std deviation
μ_k = mean

- Assumptions:
  1. For all K classes $1, 2, 3 \ldots K$ standard dev. (σ) is same
  $\therefore \sigma_1 = \sigma_2 \ldots = \sigma_K = \sigma$

Using Bayes Theorem:

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

$$= \frac{\pi_k \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \right)}{\sum_{l=1}^{K} \pi_l \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2} \right)}$$

$$\delta_k(x) = \log \pi_k - \log(\sqrt{2\pi}\sigma) + \log\left(e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}\right)$$

$$= \log \pi_k - \frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2$$

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

$$= \log \pi_k - \frac{x^2}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \frac{\mu_k x}{\sigma^2}$$

(independent of k)

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

For $K = 2$ and $\pi_1 = \pi_2$

Decision boundary: $\delta_1(x) = \delta_2(x)$

$$\Rightarrow \frac{\mu_1 x}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} = \frac{\mu_2 x}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2}$$

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

Estimating the parameters

$$\hat{\pi}_k = \frac{n_k}{n}$$

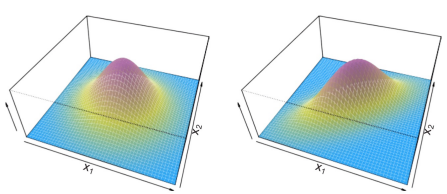
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:\, y_i = k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$$

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n - K} \cdot \hat{\sigma}_k^2$$

Where $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in the *k*th class.

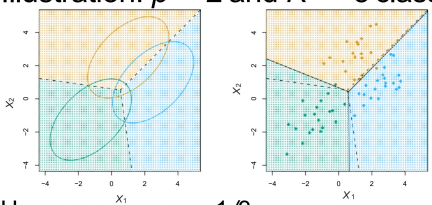
Linear Discriminant Analysis when $p > 1$



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

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

Illustration: $p = 2$ and $K = 3$ classes



- Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.
- The dashed lines are known as the *Bayes decision boundaries*. Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

Left: ellipse show 95% of the prob. of each class.

Right: 20 observations were generated from each class using LDA for decision boundary (solid line)

Practice Q.

$$f_2(x) = \frac{1}{x\sqrt{2\pi}\sigma_2} e^{\left(\frac{-(\ln x - \mu_2)^2}{2\sigma_2^2}\right)}$$

$$f_1(x) = \frac{x}{\sigma_1^2} e^{\left(\frac{-x^2}{2\sigma_1^2}\right)}$$

$$\delta_2(x) = -\log x - \frac{1}{2}\log(2\pi\sigma_2) - (\log x - \mu_2)^2 \times \frac{1}{2\sigma_2^2}$$

$$= -\log x - \frac{1}{2}\log(2\pi\sigma_2) - \frac{1}{2\sigma_2^2}((\log x)^2 + \mu_2^2 - 2\log x \cdot \mu_2)$$

$$= -\log x - \frac{1}{2}\log(2\pi\sigma_2) - \frac{(\log x)^2}{2\sigma_2^2} - \frac{\mu_2^2}{2\sigma_2^2} + \frac{\mu_2 \log x}{\sigma_2^2}$$

$$= \log x \left[\frac{\mu_2}{\sigma_2^2} - \frac{\log x}{2\sigma_2^2} - 1\right] + \log \pi_2$$

Not linear (Logarithmic)

$$\delta_1(x) = \log x - \log(\sigma_1^2) - \frac{1}{2\sigma_1^2}x^2 + \log \pi_1$$

$$= \log x - \frac{x^2}{2\sigma_1^2} + \log \pi_1$$

Not linear (Quadratic)

$\sigma_1 = \sigma_2 = 1$; $\mu_2 = 10$; $\pi_1 = \pi_2 = 0.5$; $x = 10$

$$\delta_1(x=10) = \log_e 10 - \frac{100}{2\sigma_1^2} + 0 = \text{Negative}$$

$$\delta_2(x=10) = \log_e 10 \left[10 - \frac{\log_e 10}{2} - 1\right] + 0$$

$$= 2.302\,(9.848)$$

$$= 22.67$$

$\therefore$ Class 2

$FP = \frac{23}{9667} = 0.2\%$

$FN = \frac{252}{333} = 75.67\%$

LDA on Credit Data: Confusion Matrix

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

(23 + 252)/10000 errors — a 2.75% misclassification rate!

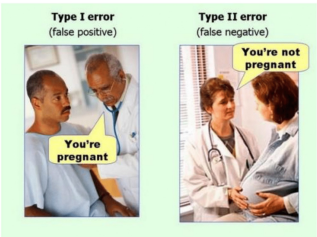
97

✡ Errors:

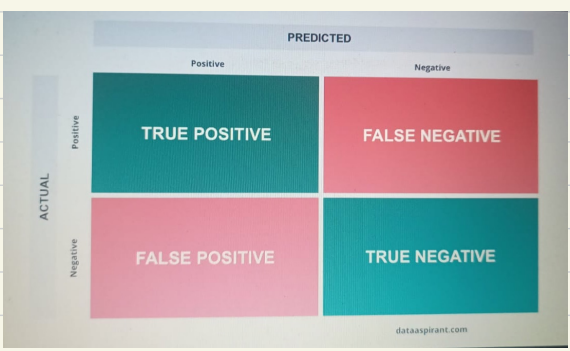
Types of errors

- False positive (type I error) rate: The fraction of negative examples that are classified as positive — 0.2% in example.
- False negative (type II error) rate: The fraction of positive examples that are classified as negative — 75.7% in example.

Types of errors



https://chemicalstatistician.files.wordpress.com/2014/05/pregnant.jpg?w=500



Event of interest = Pos.

Measures for Different Types of Error

The sensitivity or recall of a binary classifier is the rate that the event of interest is predicted correctly for all samples having the event, or

TP/P=TP/(TP+FN)

It is also called the True Positive Rate (TPR) or Hit Rate (HR).

- What proportion of credit card defaults did we detect?

Ans. $\frac{81}{333} = 24.3\%$

102

Non-event = Neg.

Measures for Different Types of Error

The specificity of a binary classifier is the rate that non-events are predicted correctly for all non-event samples or

TN/N=TN/(TN+FP)

It is also called the True Negative Rate (TNR).

- What proportion of credit card non-defaults did we detect?

Ans. $\frac{9644}{9667} = 99.7\%$

103

LDA on Credit Data: Confusion Matrix

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

(23 + 252)/10000 errors — a 2.75% misclassification rate!

97

ROC Curve

To overcome class imbalance

We produced the confusion matrix for credit data by classifying to class Yes if

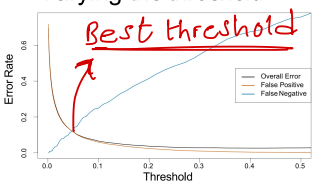
$$\widehat{\Pr}(\text{Default} = \text{Yes} \mid \text{Balance}, \text{Student}) \geq 0.5$$

We can change the two error rates by changing the threshold from 0.5 to some other value in [0, 1]:

$\widehat{\Pr}(\text{Default} = \text{Yes} \mid \text{Balance}, \text{Student}) \geq$ *threshold*, and vary *threshold*.

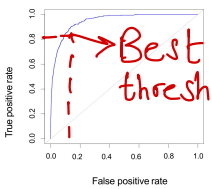
- Represents trade-off between sensitivity and specificity.
- AUC ∝ Strength of classifier
- Best trade-off ⇒ top left of curve

Varying the *threshold*



In order to reduce the false negative rate, we may want to reduce the threshold to 0.1 or less.

105



The *ROC plot* displays both simultaneously. Sometimes we use the *AUC* or *area under the curve* to summarize the overall performance. Higher *AUC* is good.

106

Measures for Different Types of Error

The precision or the positive predictive value of a binary classifier is the ratio of true positives with respected to all detected positives.

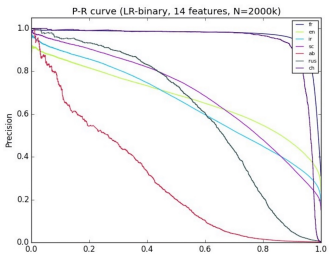
Precision= PPV=TP/(TP+FP)

Of those defaults that we detected, what proportion actually defaulted?

Ans.: $\frac{81}{104} = 77.8\%$

107

Precision-Recall Curve



From: https://stackoverflow.com/questions/33294574/good--roc--curve--but--poor--precision--recall--curve

LDA on Credit Data: Confusion Matrix

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

(23 + 252)/10000 errors — a 2.75% misclassification rate!

Measures for Different Types of Error

The negative predictive value of a binary classifier is the ratio of true negatives with respect to all detected negatives.

NPV=TN/(TN+FN)

Of those non-defaults that we detected, what proportion actually did not default?

Ans. $\frac{9644}{9896} = 97.4\%$

Harmonic Mean :

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Measures for Different Types of Error

The F1 score or F measure of a binary classifier is the harmonic mean of precision and recall:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Measures for Different Types of Error

F1 Score: seeks a balance between Precision (ratio of true positives to all detected positives) and Recall (True Positive Rate).

F1 Score might be better than accuracy if we seek a balance between Precision and Recall AND there is a class imbalance (large number of Actual Negatives). ?

Measures for Different Types of Error

F_β Score: seeks a *skewed* balance between Precision (ratio of true positives to all detected positives) and Recall (True Positive Rate).

$$F_\beta = \frac{\beta^2 + 1}{\frac{\beta^2}{recall} + \frac{1}{precision}}$$

Precision, recall, F-score is for single class.

Multiple Classes?

Two different ways of averaging:

- A **macro** average just averages the individually calculated scores of each class
  - Weights each *class* equally

$$PRE_{macro} = \frac{PRE_1 + \cdots + PRE_k}{k}$$

Multiple Classes?

Two different ways of averaging:

- A **micro** average calculates the metric by first pooling all instances of each class
  - Weights each *instance* equally

$$PRE_{micro} = \frac{TP_1 + \cdots + TP_k}{TP_1 + \cdots + TP_k + FP_1 + \cdots + FP_k}$$

Note that all of the measures used to evaluate types of error can be computed over both training and test sets.

Quadratic Discriminant Analysis :

c) **Quadratic Discriminant Analysis :**

- Unlike LDA, QDA considers each class to have its own covariance matrix.
- An observation from k^{th} class is represented as : $X \sim N(\mu_k, \Sigma_k)$ ---- Σ_k = covariance matrix of k^{th} cl
- Thus, Bayes' Classifier assigns observation $X = x$ to the class for which

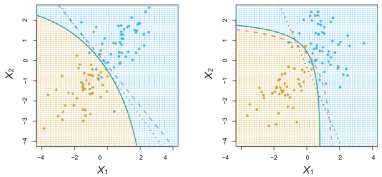
$$\delta_k(x) = \frac{-1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$$

$$= \frac{-1}{2} x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k$$

$$- \frac{1}{2}\log|\Sigma_k| + \log \pi_k$$

is largest.

Quadratic Discriminant Analysis



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

Because the Σ_k are different, the quadratic terms matter.

purple dashed = Bayes Decision Boundary
black dotted = LDA
green solid = QDA

✡ Comparison bet^n various classifiers :

Logistic Regression versus LDA

For a two-class problem, one can show that for LDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1x_1 + \ldots + c_px_p$$

So it has the same form as logistic regression.

The difference is in how the parameters are estimated.

- Logistic regression uses the conditional likelihood based on $\Pr(Y|X)$ (known as *discriminative learning*).
- LDA uses the full likelihood based on $\Pr(X, Y)$ (known as *generative learning*).
- Despite these differences, in practice the results are often very similar.

Logistic regression can also fit quadratic boundaries like QDA, by explicitly including quadratic terms in the model.

123

+ LDA is less flexible than QDA
+ LDA works well with less training observations.
+ QDA works well with large training set.

✡ Naive Bayes Classifier :

Naïve Bayes

Assumes features are independent in each class. Useful when p is large, and so multivariate methods like QDA and even LDA break down.

- Gaussian naïve Bayes assumes each Σ_k is diagonal:

$$\delta_k(x) \propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log \pi_k$$

- can use for *mixed* feature vectors (qualitative and quantitative). If X_j is qualitative, replace $f_{kj}(x_j)$ with probability mass function (histogram) over discrete categories.

Despite strong assumptions, naïve Bayes often produces good classification results.

bcz, so $\Sigma(x_1, x_2) = 0$

Bayesian: Classify to the highest density

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

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

- $f_k(x) = \Pr(X = x|Y = k)$ is the *density* for X in class k. Here we will use normal densities for these, separately in each class.
- $\pi_k = \Pr(Y = k)$ is the marginal or *prior* probability for class k.
  - We classify a new point according to which density is highest.
  - When the priors are different, we take them into account as well, and compare

$$\pi_k f_k(x) = \Pr(Y = k)\ \Pr(X = x|Y = k).$$

127

Naïve Bayes classifier

Assume independence among attributes x_i when class is given:

$f_k(x_1, x_2, \ldots, x_p) = f_k(x_1)\ f_k(x_2)\ \ldots\ f_k(x_n)$

Usually straightforward and practical to estimate $f_k(x_i) = \Pr(X_i = x_i|Y = k)$ for all x_i and k.

New sample is classified to $Y=k$ if $\pi_k \Pi_i f_k(x_i)$ is maximal.

130

Example :

How to estimate $f_k(x_i) = \Pr(X_i = x_i | Y = k)$ from data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Class priors:
$\pi^{\wedge}_k = N_k / N$
π (No) = 7/10
π (Yes) = 3/10

For discrete attributes:
$\Pr^{\wedge}$ ($X_i = x_i$ | Y=k) = | x_{ik} | / N_k
where | x_{ik} | is number of instances in class k having attribute value x_i
Examples:
$\Pr^{\wedge}$(Status = Married | No) = 4/7
$\Pr^{\wedge}$ (Refund = Yes | Yes) = 0

131

example: (Understand Carefully)

How to estimate $f_k(x_i)$ from data?

For continuous attributes:
- Discretize the range into bins
  - replace with an ordinal attribute
- Two-way split: (x_i < v) or (x_i > v)
  - replace with a binary attribute
- Probability density estimation:
  - assume attribute follows some standard parametric probability distribution (usually a Gaussian)
  - use data to estimate parameters of distribution (e.g. mean and variance)
  - once distribution is known, can use it to estimate the conditional probability $\Pr(X_i = x_i | Y = k)$

132

How to estimate $f_k(x_i)$ from data?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Gaussian distribution:

$f_k(x_i) = \Pr(X_i = x_i | Y = k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{\frac{-(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}}$

one for each (x_i, k) pair

For (Income | Class = No):
sample mean = 110
sample variance = 2975

$$\Pr(\text{Income} = 120 | \text{No}) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

133

Naïve Bayes classifier

Problem: if one of the conditional probabilities is zero, then the entire expression becomes zero.
This is a significant practical problem, especially when training samples are limited.
Ways to improve probability estimation:

$$\text{Original}: p(x_j | C_i) = \frac{N_{ji}}{N_i}$$

$$\text{Laplace}: p(x_j | C_i) = \frac{N_{ji} + 1}{N_i + c}$$

$$\text{m-estimate}: p(x_j | C_i) = \frac{N_{ji} + mp}{N_i + m}$$

c: number of levels variable x_i can take.
p: prior probability
m: parameter

134

Summary of Naïve Bayes

- Robust to isolated noise samples.
- Handles missing values by ignoring the sample during probability estimate calculations.
- Robust to irrelevant attributes.
- *NOT* robust to redundant attributes.
  - Independence assumption does not hold in this case.
  - Use other techniques such as Bayesian Belief Networks (BBN).

135

Text Classification

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …

Q. How to classify text ?

Ans: ① Hand coded rules

eg.: Spam : ① Black-list addr.

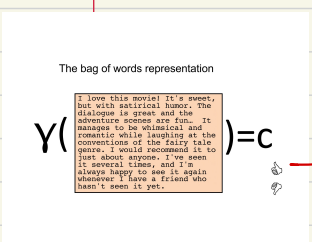
② "dollars" OR "selected"

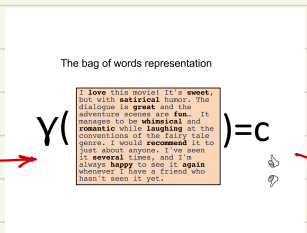
Accuracy can be improved if rules are defined carefully.

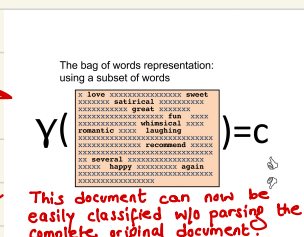
② Supervised ML algorithms :

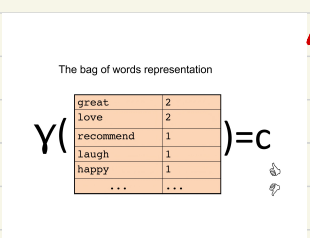
Any kind of classifier
- Naïve Bayes
- Logistic regression
- Support-vector machines
- k-Nearest Neighbors
- …

✡ Bag of words method :









+ Bag of Words is used in NLP
+ A 'bag of words is formed and each word is assigned a value (importance value)
+ For eg.: A BoW for 'good' words will be used to rate a movie.
+ Usually value is assigned according to the frequency

Application

The Bag-of-words model is mainly used as a tool of feature generation.
After transforming the text into a "bag of words", we calculate various measures to characterize the text.

The most common type of characteristics, or features calculated from the Bag-of-words model is term frequency, namely, the number of times a term appears in the text.

Example

(1) John likes to watch movies. Mary likes movies too.
John also likes to watch football games.

Combined Documents:

```
BoW3 =
{"John":2,"likes":3,"to":2,"watch":2,"movies":2,"Mary":1,"too":1,"also":1,"football":1,"games":1};
```

Term Frequency (TF)

The simplest choice to calculate Term Frequency to measure the importance of a word in a document is to use the *raw count* of a term in a document.

tf(t,d)=number of occurrences of term t in document d

Other choices:
- Boolean Frequencies tf(t,d)=1 if t occurs in d otherwise 0
- Adjustment for document length: tf(t,d)=[number of occurrences of term t in document d]/number of words in document d

Inverse Document Frequency (IDF)

A measure of information each word provides, i.e. if a word is common or rare among documents

IDF(t,D) = log(total number of documents in the corpus D/number of documents where the term t appears)

TFIDF

A combined measure for each word is the TF-IDF which is a product of TF and IDF: TFIDF = TF x IDF

TFIDF(t,d,D)=TF(t,d).IDF(t,D)

Therefore, each document can be represented as a vector representing the bag of words, but instead of simple frequencies, the TFIDF of each word can be given in the vector.