

FIT2086 Lecture 7

Classification, Naïve Bayes and logistic regression

Daniel F. Schmidt

Faculty of Information Technology, Monash University

September 4, 2017

Outline

- 1 Classifiers
 - Naïve Bayes Classifiers
 - Logistic Regression

- 2 Assessing Classifiers
 - How good is a classifier?

Revision from last week (1)

- Linear regression

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i}$$

- β_0 is the **intercept** (value of $\mathbb{E}[Y]$ when all predictors are zero)
 - β_j is a **coefficient** (change in $\mathbb{E}[Y]$ per unit change in $x_{j,i}$)
- Residuals (errors)

$$e_i = y_i - \beta_0 - \beta_1 x_{1,i} - \beta_2 x_{2,i} - \cdots - \beta_p x_{p,i}$$

- Residual sum-of-squares

$$\text{RSS}(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n e_i^2$$

- Least-squares estimates linear model by finding $\beta_0, \beta_1, \dots, \beta_p$ that minimise the RSS

Revision from last week (2)

- R^2 goodness-of-fit

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where TSS is the sum of squared errors for the mean model

- Model fitting:
 - Overfitting = including unimportant predictors
 - Underfitting = excluding important predictors
- Hypothesis testing to determine if variable is important
 - Test $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$
 - The smaller the p -value the stronger predictor j is associated with the target
- Model selection methods:
 - Add complexity penalty to the negative log-likelihood
- Finding good models – all subsets selection, stepwise selection

Outline

- 1 Classifiers
 - Naïve Bayes Classifiers
 - Logistic Regression
- 2 Assessing Classifiers
 - How good is a classifier?

Supervised Learning – recap (1)

- Imagine we have measured $p + 1$ variables on n individuals (people, objects, things)
- We would like to predict one of the variables using the remaining p variables
- If the variable we are predicting is categorical, we are performing **classification**
 - Example: predicting if someone has diabetes from medical measurements.
- If the variable we are predicting is numerical, we are performing **regression**
 - Example: Predicting the quality of a wine from chemical and seasonal information.

Supervised Learning – recap (1)

- Imagine we have measured $p + 1$ variables on n individuals (people, objects, things)
- We would like to predict one of the variables using the remaining p variables
- If the variable we are predicting is categorical, we are performing **classification**
 - Example: predicting if someone has diabetes from medical measurements.
- If the variable we are predicting is numerical, we are performing **regression**
 - Example: Predicting the quality of a wine from chemical and seasonal information.

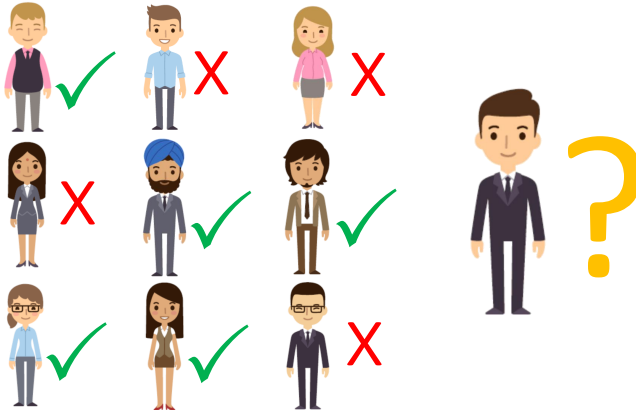
Supervised Learning – recap (2)

- The variable we are predicting is designated the “y” variable
 - We have (y_1, \dots, y_n)
- This variable is often called the:
 - target;
 - response;
 - outcome.
- The other variables are usually designated “X” variables
 - We have $(x_{i,1}, \dots, x_{i,n})$ for $i = 1, \dots, p$
- These variables are often called the
 - explanatory variables;
 - predictors;
 - covariates;
 - features;
 - exposures.
- Usually we assume the targets are random variables and the predictors are known without error

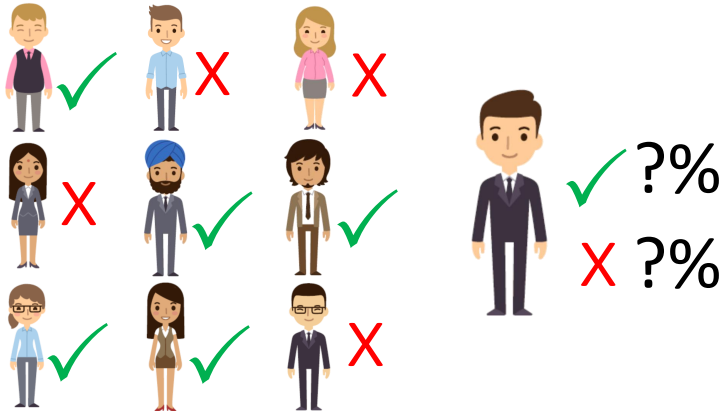
Supervised Learning – recap (2)

- The variable we are predicting is designated the “y” variable
 - We have (y_1, \dots, y_n)
- This variable is often called the:
 - target;
 - response;
 - outcome.
- The other variables are usually designated “X” variables
 - We have $(x_{i,1}, \dots, x_{i,n})$ for $i = 1, \dots, p$
- These variables are often called the
 - explanatory variables;
 - predictors;
 - covariates;
 - features;
 - exposures.
- Usually we assume the targets are random variables and the predictors are known without error

Classifiers



Probabilistic classifiers



Classification

- We begin by defining a classifier in terms of probability
- Imagine we have a **categorical** outcome variable Y
- We also have p predictor variables X_1, \dots, X_p
 \Rightarrow often called **features** in classification literature
- We can build a classifier for Y using our predictors, i.e., we want to find

$$\mathbb{P}(Y = y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p)$$

- This is the **conditional probability** of Y given X_1, \dots, X_p .
- Gives us the probability of an individual being in class $Y = y$, given the values of their predictors x_1, \dots, x_p

Classifiers (1)

- Let us now specialise our problem
- Assume that all the predictors are also **categorical**
- The formula for conditional probability is:

$$\mathbb{P}(Y = y \mid X_1 = x_1, \dots, X_p = x_p) = \frac{\mathbb{P}(Y = y, X_1 = x_1, \dots, X_p = x_p)}{\mathbb{P}(X_1 = x_1, \dots, X_p = x_p)}$$

where

- the numerator is the **joint probability** of $(Y = y, X_1 = x_1, \dots, X_p = x_p)$;
- the denominator is the **marginal probability** of $(X_1 = x_1, \dots, X_p = x_p)$.

Classifiers (2)

- So if we have the **joint probability** we can build a classifier
- Recall example from Assignment 1:

	No Heart Disease ($H = 0$)	Heart Disease ($H = 1$)
No Mutation ($M = 0$)	0.35	0.30
Mutation ($M = 1$)	0.10	0.25

Population joint probabilities of heart disease/LDLR mutation.

- Then we had

$$P(H = 1 | M = 0) = \frac{P(H = 1, M = 0)}{P(M = 0)} = 0.4615$$

$$P(H = 1 | M = 1) = \frac{P(H = 1, M = 1)}{P(M = 1)} = 0.7143$$

Classifiers (3)

- In our example we got told the population joint probabilities
- But in reality we don't know these – we just have data
- We can try and estimate them from the data
- For our example:
 - our target is heart disease, $H \in \{0, 1\}$,
 - predictor is LDLR mutation, $M \in \{0, 1\}$

Classifiers (4)

- Imagine we had n realisations of this random variables, $\mathbf{m} = (m_1, \dots, m_n)$ and $\mathbf{h} = (h_1, \dots, h_n)$
- We could estimate joint probability by proportions

$$F(H = h, M = m) = \frac{1}{n} \sum_{i=1}^n I(h_i = h \text{ and } m_i = m)$$

- Weak law of large numbers guarantees this will converge on population proportions for large enough n

Classifiers (5)

- Example, imagine we had
 - $\mathbf{m} = (1, 1, 0, 1, 1, 1, 0, 0)$ and
 - $\mathbf{h} = (1, 0, 1, 1, 0, 0, 1, 0)$
- Then estimated joint probabilities are

	No Heart Disease ($H = 0$)	Heart Disease ($H = 1$)
No Mutation ($M = 0$)	1/8	2/8
Mutation ($M = 1$)	3/8	2/8

Estimated joint probabilities of heart disease/LDLR mutation

- Now we can estimate $\mathbb{P}(H = h \mid M = m)$ using

$$\frac{F(H = h, M = m)}{F(H = 0, M = m) + F(H = 1, M = m)}$$

- For large n the proportions will be close to population probabilities

Classifiers (6)

- Simple enough – but there is a problem
- For our simple problem H and M were binary
⇒ only need to estimate $2 \times 2 = 4$ joint probabilities
- What if we had two binary genetic mutations, M_1 and M_2 ?
- Now have $2 \times 2 \times 2 = 8$ probabilities to estimate
- For p predictors, there are 2^{p+1} probabilities to estimate
⇒ **exponential growth** in p
- This rapidly outstrips our sample size n no matter how big n is

Classifiers (7)

- We need to constrain the problem
- Two simple approaches popular in literature
- **Naïve Bayes** classifiers
 - Make simplifying assumptions about joint probabilities
 - Easily handle categorical predictors
 - Easily handles multi-class targets
 - Popular in text mining and classification
- **Logistic regression**
 - Adaptation of the linear model, widely used
 - Directly estimates conditional probabilities
 - Handles categorical and continuous predictors
 - More difficult to handle multi-class targets

Naïve Bayes Classifiers (1)

- The Naïve Bayes solves the problem of too many probabilities to estimate by making a very strong assumption
- Let X_1, \dots, X_p be p categorical predictors (features)
 \Rightarrow do not have to be binary
- Use the shorthand notation $\mathbb{P}(Y = y, X = x) \equiv P(y, x)$
- Naïve Bayes assumes predictors are **conditionally independent**, given the value of the target

$$P(x_1, \dots, x_p | y) = \prod_{j=1}^p P(x_j | y)$$

Naïve Bayes Classifiers (2)

- This assumption lets us write joint probability as

$$\begin{aligned}P(y, x_1, \dots, x_p) &= P(x_1, \dots, x_p | y)P(y) \\&= \prod_{j=1}^p P(x_j | y)P(y)\end{aligned}$$

- Using this joint probability we can write

$$P(y | x_1, \dots, x_p) = \frac{\prod_{j=1}^p P(x_j | y)P(y)}{P(x_1, \dots, x_p)}$$

where we note $P(x_1, \dots, x_p)$ does not depend on y
 \Rightarrow constant for given values of predictors

Naïve Bayes Classifiers (3)

- The formula

$$P(y | x_1, \dots, x_p) = \frac{\prod_{j=1}^p P(x_j | y) P(y)}{P(x_1, \dots, x_p)}$$

is a statement of **Bayes'** theorem.

- You can think of it this way:
 - $P(x_j | y)$ is the probability of seeing a value for predictor, given that the target has a particular value (**likelihood** of x_j)
 - $P(y)$ is the probability of the target having a particular value in our population (**prior** of y)
- Bayes' rule takes the prior probability of seeing $Y = y$, and updates this probability to take into account the fact that we have also observed the features x_1, \dots, x_p .
- The resulting probability is called the **posterior probability**.

Naïve Bayes Classifiers (3)

- The formula

$$P(y | x_1, \dots, x_p) = \frac{\prod_{j=1}^p P(x_j | y) P(y)}{P(x_1, \dots, x_p)}$$

is a statement of **Bayes'** theorem.

- You can think of it this way:
 - $P(x_j | y)$ is the probability of seeing a value for predictor, given that the target has a particular value (**likelihood** of x_j)
 - $P(y)$ is the probability of the target having a particular value in our population (**prior** of y)
- Bayes' rule takes the prior probability of seeing $Y = y$, and updates this probability to take into account the fact that we have also observed the features x_1, \dots, x_p .
- The resulting probability is called the **posterior probability**.

Naïve Bayes Classifiers (4)

- Reduced number of probabilities we need to estimate
 - Let predictor j be categorical with K_j categories
 - Let target Y be categorical with K_y categories
- Conditional independence means we need only estimate

$$\mathbb{P}(X_j = x \mid Y = y), \quad x \in \{1, \dots, K_j\}, \quad y \in \{1, \dots, K_y\}$$

for $j = 1, \dots, p$, and

$$\mathbb{P}(Y = y), \quad y \in \{1, \dots, K_y\}$$

Naïve Bayes Classifiers (5)

- Without assumptions we estimate $K_y \prod_{j=1}^p K_j$ probabilities
- With assumptions we only estimate

$$K_y \times (K_y \times K_1) + (K_y \times K_2) + \dots + (K_y \times K_p)$$

probabilities, and

$$K_y^{p+1} \sum_{j=1}^p K_j \ll K_y \prod_{j=1}^p K_j$$

- So Naïve Bayes assumptions reduce problem size to **linear** in p

Naïve Bayes Classifiers (6)

- Given n samples of target-predictor pairs

- $\mathbf{y} = (y_1, \dots, y_n)$
- $\mathbf{x}_j = (x_{1,j}, \dots, x_{n,j}), j = 1, \dots, p$

we can learn a Naïve Bayes model from the proportions:

$$F(X_j = x | Y = y) = \frac{\sum_{i=1}^n I(x_{i,j} = x \text{ and } y_i = y)}{\sum_{i=1}^n I(y_i = y)}$$

and

$$F(Y = y) = \frac{1}{n} \sum_{i=1}^n I(y_i = y)$$

- Learning a Naïve Bayes classifier is linear time complexity in the sample size n .

Naïve Bayes Classifiers (7)

- Given the proportions we can classify a new set of predictor values x'_1, \dots, x'_p using

$$P(y | x'_1, \dots, x'_p) = \frac{\prod_{j=1}^p F(X_j = x'_j | Y = y) F(Y = y)}{\sum_{y \in \{1, \dots, K_y\}} \prod_{j=1}^p F(X_j = x'_j | Y = y) F(Y = y)}$$

which is linear in the feature size p .

- So Naïve Bayes is **efficient**, but only by being quite **restrictive**
- No real reason to think that predictors should be conditionally independent in real life
- However, often works sufficiently well in practice

Naïve Bayes Classifiers (8)

- What if we want to guess which class an individual is in?
⇒ Find the value of y which maximises $P(y | x'_1, \dots, x'_p)$
- Given predictor values x'_1, \dots, x'_p

$$P(y | x'_1, \dots, x'_p) \propto \prod_{j=1}^p F(X_j = x'_j | Y = y) F(Y = y)$$

as denominator on right-hand-side is constant.

Naïve Bayes Classifiers (9)

- So far assumed categorical predictors
- How to handle continuous predictors?
- There are two approaches:

- 1 We assume parametric distributions for each predictor, i.e.,

$$p(x_j | y) \equiv p(x_j | \theta_{j,y})$$

and estimate $\hat{\theta}_{j,y}$ from the data using maximum likelihood

- 2 We discretize the predictor into categories, i.e., use histogram
- We do not examine this further in this subject

Logistic Regression (1)

- The Naïve Bayes approach made assumptions about the features to simplify estimation of the **joint probabilities**
- It then used the joint probabilities to find the conditional probability of the targets using Bayes rule.
- This is a round-about way of solving problem
- Logistic regression directly models the conditional probabilities
 - Extends the **linear regression** model to binary data

Logistic Regression (2)

- Given predictors $x_{i,1}, \dots, x_{i,p}$ multiple linear regression predicts the target as

$$\mathbb{E}[y_i] = \eta_j = \beta_0 + \sum_{j=1}^p \beta_j x_{i,j}$$

where η_i is shorthand for our **linear predictor** for individual i

- We find $\beta_0, \beta_1, \dots, \beta_j$ by least-squares
- If our target is binary, we *could* fit a linear model using least-squares and approximate

$$\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p}) \approx \eta_i$$

- Serious problem: our predicted value η_i could be less than zero, or greater than one, for certain values of the features!

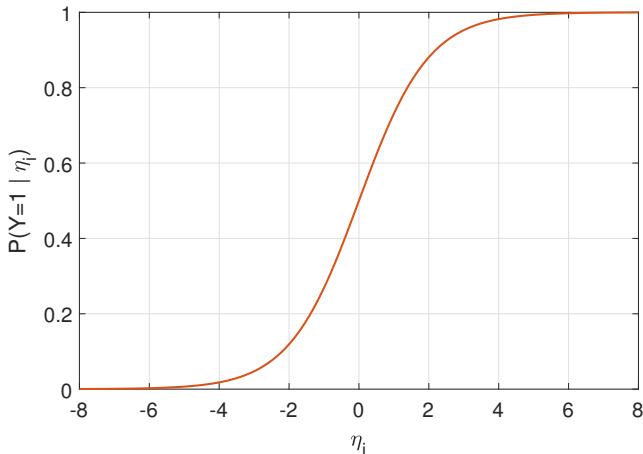
Logistic Regression (3)

- One solution is to bound η_i to $(0,1)$
- There exist a lot of ways of bounding η_i
- Logistic regression chooses to use the **logistic function**

$$\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p}) = \frac{1}{1 + \exp(-\eta_i)}$$

- This function smoothly
 - tends to 0 as $\eta_i \equiv \beta_0 + \sum_{j=1}^p \beta_j x_{i,j} \rightarrow -\infty$;
 - tends to 1 as $\eta_i \equiv \beta_0 + \sum_{j=1}^p \beta_j x_{i,j} \rightarrow \infty$.

Logistic Regression (4)



The logistic function. As $\eta_i \rightarrow -\infty$, then $P(Y = 1) \rightarrow 0$, and as $\eta_i \rightarrow \infty$, then $P(Y = 1) \rightarrow 1$.

Logistic Regression (5)

- We can interpret the linear model in terms of **log-odds**
- Given the $\mathbb{P}(Y = 1)$ and $\mathbb{P}(Y = 0)$, the odds for $Y = 1$ are

$$\mathbb{P}(Y = 1)/\mathbb{P}(Y = 0)$$

- They reflect how many more times likely the event $Y = 1$ is to occur than the event $Y = 0$
- **Example:** probability of a heads from coin toss is 0.75; then
 - the odds for seeing a head are $0.75/0.25 = 3$;
 - the odds for seeing a tail are $0.25/0.75 = 1/3$.
- The log-odds make this symmetric:
 - the log-odds for seeing a head are $\log(0.75/0.25) = \log 3$;
 - the log-odds for seeing a tail are $\log(0.25/0.75) = -\log 3$.

Logistic Regression (5)

- We can interpret the linear model in terms of **log-odds**
- Given the $\mathbb{P}(Y = 1)$ and $\mathbb{P}(Y = 0)$, the odds for $Y = 1$ are

$$\mathbb{P}(Y = 1)/\mathbb{P}(Y = 0)$$

- They reflect how many more times likely the event $Y = 1$ is to occur than the event $Y = 0$
- **Example:** probability of a heads from coin toss is 0.75; then
 - the odds for seeing a head are $0.75/0.25 = 3$;
 - the odds for seeing a tail are $0.25/0.75 = 1/3$.
- The log-odds make this symmetric:
 - the log-odds for seeing a head are $\log(0.75/0.25) = \log 3$;
 - the log-odds for seeing a tail are $\log(0.25/0.75) = -\log 3$.

Logistic Regression (6)

- A logistic regression models the conditional log-odds as

$$\log \left(\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{\mathbb{P}(Y_i = 0 \mid x_{i,1}, \dots, x_{i,p})} \right) = \beta_0 + \sum_{j=1}^p \beta_j x_{i,j} \equiv \eta_i$$

- So the log-odds of a success, given the values of the predictors, is equal to the linear predictor η_i
 - the **intercept** β_0 is the log-odds when all the predictors are zero, i.e., $x_{i,1} = x_{i,2} = \dots = x_{i,p} = 0$;
 - the **coefficient** β_j is the increase in log-odds per unit change of predictor x_j
- The odds for $Y = 1$ are $\exp(\eta_i)$
 - when $\eta_j > 0$, $Y = 1$ is more likely than $Y = 0$;
 - when $\eta_j < 0$, $Y = 0$ is more likely than $Y = 1$

Logistic Regression (7)

- To see that setting log-odds equal to η_i leads to logistic regression, write:

$$\log \left(\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{1 - \mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})} \right) = \eta_i$$

- Now exponentiate both sides

$$\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{1 - \mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})} = \exp(\eta_i)$$

- Solving for $\mathbb{P}(Y_i = 1 \mid \dots)$ yields

$$\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p}) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)}$$

(noting $1/e^a = e^{-a}$) which is the logistic function.

Logistic Regression (7)

- To see that setting log-odds equal to η_i leads to logistic regression, write:

$$\log \left(\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{1 - \mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})} \right) = \eta_i$$

- Now exponentiate both sides

$$\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{1 - \mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})} = \exp(\eta_i)$$

- Solving for $\mathbb{P}(Y_i = 1 \mid \dots)$ yields

$$\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p}) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)}$$

(noting $1/e^a = e^{-a}$) which is the logistic function.

Logistic Regression (7)

- To see that setting log-odds equal to η_i leads to logistic regression, write:

$$\log \left(\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{1 - \mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})} \right) = \eta_i$$

- Now exponentiate both sides

$$\frac{\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})}{1 - \mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p})} = \exp(\eta_i)$$

- Solving for $\mathbb{P}(Y_i = 1 \mid \dots)$ yields

$$\mathbb{P}(Y_i = 1 \mid x_{i,1}, \dots, x_{i,p}) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} = \frac{1}{1 + \exp(-\eta_i)}$$

(noting $1/e^a = e^{-a}$) which is the logistic function.

Fitting Logistic Regression Models (1)

- How to estimate the regression coefficients?
- Many packages use maximum likelihood
- Let $\beta = (\beta_1, \dots, \beta_p)$ be our regression coefficients
- Assume our targets are independent RVs;
- Each Y_i is then distributed as per a Bernoulli distribution

$$Y_i \sim \text{Be}(\theta_i(\beta_0, \beta))$$

where

$$\theta_i(\beta_0, \beta) = \frac{1}{1 + \exp\left(-\beta_0 - \sum_{j=1}^p \beta_j x_{i,j}\right)}$$

is the probability of success for individual i , given the predictors $x_{i,1}, \dots, x_{i,p}$ and the parameters $\beta_0, \beta_1, \dots, \beta_p$.

Fitting Logistic Regression Models (2)

- If $\mathbf{y} = (y_1, \dots, y_n)$ are binary targets, the likelihood for a logistic regression is then

$$\begin{aligned} p(\mathbf{y} \mid \beta_0, \boldsymbol{\beta}) &= \prod_{i=1}^n p(y_i \mid \beta_0, \boldsymbol{\beta}) \\ &= \prod_{i=1}^n \theta_i(\beta_0, \boldsymbol{\beta})^{y_i} (1 - \theta_i(\beta_0, \boldsymbol{\beta}))^{1-y_i} \end{aligned}$$

from the pdf of the Bernoulli distribution.

- The negative log-likelihood is then

$$L(\mathbf{y} \mid \beta_0, \boldsymbol{\beta}) = - \sum_{i=1}^n [y_i \log \theta_i(\beta_0, \boldsymbol{\beta}) + (1 - y_i) \log (1 - \theta_i(\beta_0, \boldsymbol{\beta}))]$$

Fitting Logistic Regression Models (3)

- The negative log-likelihood is then

$$L(\mathbf{y} \mid \beta_0, \boldsymbol{\beta}) = - \sum_{i=1}^n [y_i \log \theta_i(\beta_0, \boldsymbol{\beta}) + (1 - y_i) \log (1 - \theta_i(\beta_0, \boldsymbol{\beta}))]$$

- The values $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$ of β_0 and $\boldsymbol{\beta}$ that minimise this quantity are the maximum likelihood estimates
- No closed form solution exists, must be found numerically
⇒ But luckily always only a single, global minimum
- Time complexity roughly cubic in number of predictors p
⇒ potentially slower than Naïve Bayes
- But much more interpretable

Goodness-of-fit

- The minimised negative log-likelihood

$$L(\mathbf{y} \mid \hat{\beta}_0, \beta)$$

is a measure of goodness-of-fit of a model.

- The difference in minimised negative log-likelihoods

$$L(\mathbf{y} \mid \hat{\beta}_0) - L(\mathbf{y} \mid \hat{\beta}_0, \hat{\beta})$$

is a measure similar to R^2 ; (bigger differences \Rightarrow better fit)

- Relative to model with an intercept only (no predictors); equivalent to assuming $P(Y = 1)$ is same for all individuals
- Sometimes maximised log-likelihood is reported instead, or two times (negative) log-likelihoods; depends on package

Predicting with a Logistic Regression

- Once we have found estimates $\hat{\beta}_0, \hat{\beta}$ it is easy to predict with a logistic regression
- For some new values of features x'_1, \dots, x'_p , we calculate

$$\hat{\eta} = \hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j x'_j$$

- Then we can estimate the probability that $Y' = 1$ for these features:

$$\mathbb{P}(Y' = 1 \mid x'_1, \dots, x'_p) = \frac{1}{1 + \exp(-\hat{\eta})}$$

- If we need to guess at most likely class, choose value of Y' that maximises this probability

Extensions

- A strength of logistic regression is that it builds on the tools used in linear regression
- Handle categorical predictors same as linear regression
⇒ form new indicator variables for each category
- They can also handle non-linearities the same way as linear regressions, e.g.,
 - logarithmic transformations of predictors;
 - polynomial transformations of predictors.

Finding logistic regression models

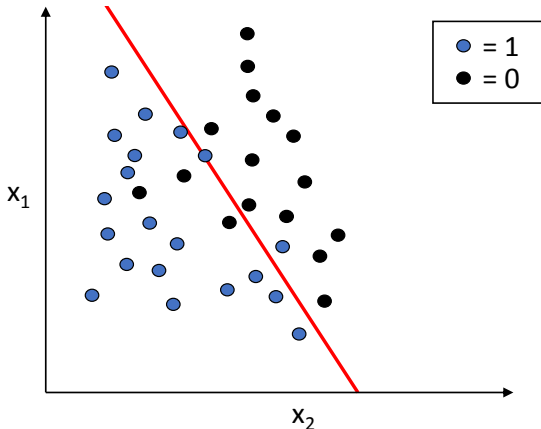
- Same approaches as for linear models
⇒ try to avoid under/over-fitting
- Hypothesis testing $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$
 - Smaller p -value \Rightarrow more likely predictor j is important
- Model selection
 - Use penalized likelihood:

$$L(\mathbf{y} | \hat{\beta}_0, \hat{\beta}) + k\alpha_n$$

where:

- k is number of predictors in model;
 - $\alpha_n = 1$ for Akaike information criterion (AIC);
 - $\alpha_n = 3/2$ for Kullback information criterion (KIC);
 - $\alpha_n = (1/2) \log n$ for Bayesian information criterion (BIC).
- Sometimes two times these quantities are used (e.g., in R)
- Forward/backwards selection of predictors

Logistic regressions are linear



A logistic regression separates successes from failures by using a linear separation surface. The line is defined by the values of the two features x_1 and x_2 that satisfy $\mathbb{P}(Y = 1 | x_1, x_2) = 1/2$. For models with p features, this becomes a p -dimensional plane.

Outline

- 1 Classifiers
 - Naïve Bayes Classifiers
 - Logistic Regression
- 2 Assessing Classifiers
 - How good is a classifier?

Performance Measures for Classifiers (1)

- Imagine we have trained a classifier on some data (logistic regression, Naïve Bayes, for something else)
- We now get a new body of data and want to test how well our classifier performs
- What measures of performance exist for classification problems?
 - Classification error
 - Sensitivity and specificity
 - Area-under-the-curve (AUC)
 - Logarithmic loss

Performance Measures for Classifiers (2)

- Let $\mathbf{y}' = (y'_1, \dots, y'_{n'})$ be a vector of new data to test on
 - For simplicity, let us assume y'_i is binary
- Let $\mathbf{x}'_j = (x'_{1,j}, \dots, x'_{n',j})$ be the vector for feature j
- For each of the new individuals, we can calculate our best guess at which class it belongs to using:

$$\hat{y}'_i = \arg \max_{y \in \{0,1\}} \left\{ \mathbb{P}(Y'_i = y \mid x'_{i,1}, \dots, x'_{i,p}) \right\}$$

where the probabilities are estimated using the model we have learned from our training data

Classification Accuracy (1)

- The most straightforward measure of performance is **classification accuracy**
- This is given by:

$$\text{CA} = \frac{1}{n'} \sum_{i=1}^{n'} I(y'_i = \hat{y}'_i)$$

where $I(\cdot)$ is one if the condition inside the parenthesis is met, and a zero otherwise

- The proportion of times our classifier correctly guesses the class of a new individual
- This ranges from 0 (perfectly incorrect), through to 1/2 (only as good as random guessing) to 1 (perfectly correct)
- Realistically 1/2 is worst accuracy – if $< 1/2$ we can swap our classification output

Classification Accuracy (2)

- More generally, we can form a **confusion matrix**

	$y_i = 0$	$y_i = 1$
$\hat{y}_i = 0$	True Negative (TN)	False Negative (FN)
$\hat{y}_i = 1$	False Positive (FP)	True Positive (TP)

- Classification accuracy is then

$$CA = \frac{TP + TN}{TP + TN + FP + FN}$$

Specificity and Sensitivity (1)

- Can form other useful information
- **Specificity** is the true positive rate:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **Sensitivity** is the true negative rate:

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- Sensitivity of 1 means we correctly classify all individuals for which $y'_i = 1$
 - Specificity of 1 means we correctly classify all individuals for which $y'_i = 0$
- High sensitivity can be achieved at expense of decreased specificity, and vice versa

Specificity and Sensitivity (2)

- Set a detection threshold $T \in (0, 1)$ for our classifier.
- For each of the new individuals, we say that $\hat{y}'_i = 1$ if

$$\mathbb{P}(Y'_i = 1 \mid x'_{i,1}, \dots, x'_{i,p}) \geq T,$$

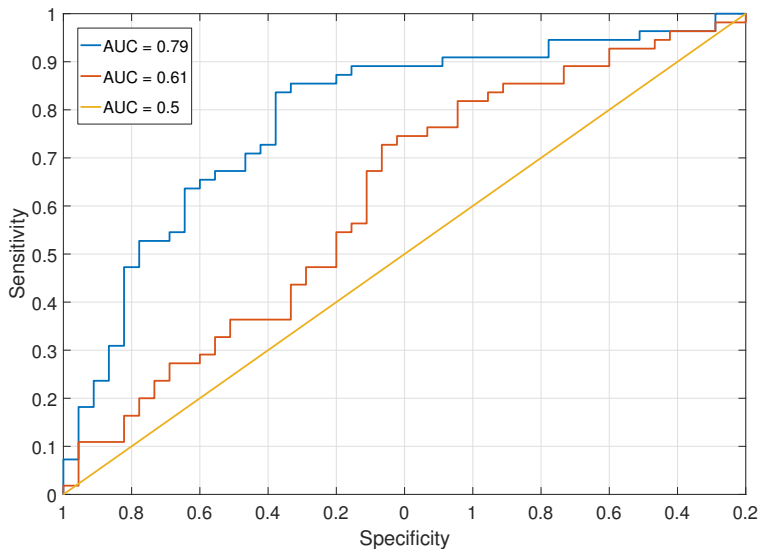
otherwise they are in class $\hat{y}'_i = 0$.

- Classifying an individual based on which class is more likely is equivalent to using $T = 1/2$
- Varying T produces different values of sensitivity, specificity
- Let $\text{TPR}(T)$, $\text{TNR}(T)$ be the sensitivity, specificity for threshold T
 - Small $T \Rightarrow$ increased sensitivity;
 - Large $T \Rightarrow$ increased specificity

Area-under-the-curve (AUC) (1)

- If we vary T from 0 to 1 we get a range of different classification rules
- Each has a different TPR and TPN
- We can plot these different values to get a “receiver operating curve”
- The area under this curve is called the **AUC**
- The bigger the area, the better the classifier

Area-under-the-curve (AUC) (2)



Area-under-the-curve (AUC) (2)

- AUC is always between 0 and 1
 - AUC of 1 is perfect classification;
 - AUC of 0 is perfect misclassification;
 - AUC of $1/2$ means we do no better than a random guess.
- How to interpret AUC?
- An AUC of p means that if we randomly sampled an individual i from our test group for whom $y'_i = 1$, and randomly sampled an individual k from our test group for whom $y'_k = 0$, then the AUC is the probability that

$$\mathbb{P}(Y'_i = 1 \mid x'_{i,1}, \dots, x'_{i,1}) > \mathbb{P}(Y'_k = 1 \mid x'_{k,1}, \dots, x'_{k,1})$$

that is, it is the probability that sample i is rated more likely to be in class 1 than sample k .

Logarithmic Loss (1)

- The final performance measure we consider is logarithmic loss
- For each sample i in our test group, we score

$$L(y'_i) = \begin{cases} \log \mathbb{P}(Y'_i = 1 \mid x'_{i,1}, \dots, x'_{i,1}) & \text{for } y'_i = 1 \\ 1 - \log \mathbb{P}(Y'_i = 1 \mid x'_{i,1}, \dots, x'_{i,1}) & \text{for } y'_i = 0 \end{cases}.$$

This is the negative-log-probability of the test data point y'_i under our classification model

- The total logarithmic loss is then

$$L(\mathbf{y}) = \sum_{i=1}^n L(y'_i)$$

- The negative-log-likelihood of this new, future data
- Smaller the score, the better our classifier predicts this data
- Log-loss measures how well the model predicts the **probabilities** of an individual being in a class

Logarithmic Loss (2)

- Classification accuracy measures how good our guesses at most likely class are
- Log-loss measures how well the model predicts the **probabilities** of an individual being in a class
- Why is this important?
- It tells you how confident you should be in your predicted class
- **Example:** Both $P(Y = 1 | x_1, \dots, x_p) = 0.501$ and $P(Y = 1 | x_1, \dots, x_p) = 0.99$ would predict the most likely class for Y to be $Y = 1$.
- We are much more confident about the latter than the former.
- Estimating conditional probabilities well lets us get better idea of how confident we should be in our predicted classes

Reading/Terms to Revise

- Reading for this week: Chapter 9 of Ross.
- Terms you should know:
 - Conditional independence
 - Bayes rule, Naïve Bayes classifier
 - Odds, log-odds
 - Logistic regression
 - Classification accuracy
 - Specificity, sensitivity
 - Area-under-the-curve (AUC)
 - Logarithmic loss
- Next week we will be examined some more recent developments in fitting and estimating linear and logistic regression models.