QBUS3820: Statistical Learning and Data Mining

Lecture 5: Review of Estimation Methods and Introduction to Classification

Semester 1, 2019

Discipline of Business Analytics, The University of Sydney Business School

Lecture 5: Review of Estimation Methods and Intro to Classification

1. Empirical risk minimisation
2. Maximum Likelihood Estimation (MLE)
3. Bayesian approach
4. Introduction to classification
5. Introduction to decision theory for classification
6. K-nearest neighbours classifier

Empirical risk minimisation

Empirical risk minimisation

Let $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ be the training data and let $f(\boldsymbol{x}; \boldsymbol{\theta})$ denote the candidate prediction functions, which depend on the parameter $\boldsymbol{\theta}$.

Recall: the **empirical risk minimisation** solves the following optimisation problem, in which L is the loss function:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L\Big(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta})\Big)$$

The $\underset{\boldsymbol{\theta}}{\operatorname{argmin}}$ operation identifies the value of θ that minimises the function on the right-hand side.

Note that in the case of the squared error loss this approach corresponds to the least squares estimation.

Regularised empirical risk minimisation

Minimising the empirical risk will typically lead to overfitting. In **regularised** empirical risk minimisation we solve:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} L\Big(y_i, f(\boldsymbol{x}_i; \boldsymbol{\theta})\Big) + \lambda\, C(\boldsymbol{\theta})\right]$$

where $C(\boldsymbol{\theta})$ is some measure of the complexity of the prediction function, and λ is a non-negative weight in the complexity penalty $\lambda\, C(\boldsymbol{\theta})$.

Maximum Likelihood Estimation (MLE)

Before formally defining MLE, we will consider a brief introductory example.

Probability vs. Statistics

Probability Question: X counts the number of successes in 20 independent random trials with probability of success π (i.e. X has Binomial distribution with parameters 20 and π).

Assume $\pi = 0.3$. What is the probability of observing $X = 4$?

$$P(X = 4) = \binom{20}{4}(0.3)^4(1 - 0.3)^{20-4} = 0.1304$$

Statistics Question: X has Binomial distribution with parameters 20 and π. We observed $X = 4$. How do we estimate π?

Statistics question

X has Binomial distribution with parameters 20 and π. We observed $X = 4$. How do we estimate π?

A simple special case:

Suppose we know that π is either 0.3 or 0.6. Which parameter value should we choose based on the observed data, $X = 4$?

$$P(X = 4) = \binom{20}{4}\pi^4(1-\pi)^{20-4}$$

$$P(X = 4\,;\,\pi = 0.3) = \binom{20}{4}(0.3)^4(0.7)^{20-4} = 0.1304$$

$$P(X = 4\,;\,\pi = 0.6) = \binom{20}{4}(0.6)^4(0.4)^{20-4} = 0.0003$$

Statistics question (special case)

Suppose we know that π is either 0.3 or 0.6. Which parameter value should we choose based on the observed data, $X = 4$?

$$P(X = 4\,;\, \pi = 0.3) = 0.1304$$

$$P(X = 4\,;\, \pi = 0.6) = 0.0003$$

Under the choice $\pi = 0.3$, the observed data, $X = 4$, is much more *likely*. Thus, it makes sense to pick $\pi = 0.3$ from the available two options.

This is the main idea of the maximum likelihood approach.

Statistics question

X has Binomial distribution with parameters 20 and π. We observed $X = 4$.

How do we estimate π?

Given the observed data, the *Likelihood* is a function of the unknown parameter:

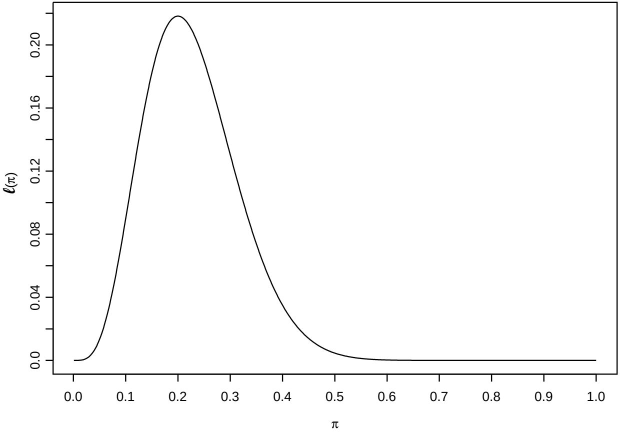
$$\ell(\pi) = P(X = 4\,;\, \pi) = \binom{20}{4}\pi^4(1-\pi)^{20-4}$$

$\ell(\pi)$ gives the probability of observing the data at hand for each value of the parameter π.

MLE: choose the value of π that maximizes $\ell(\pi)$

In other words, choose π that corresponds to the maximum probability of observing the data at hand

Likelihood function, $\ell(\pi)$, in the binomial example



MLE: $\widehat{\pi} = 0.2$

Notation

- Let $p(y; \boldsymbol{\theta})$ denote a probability mass function or a density function (for a random variable Y), which depends on the parameter $\boldsymbol{\theta}$.

- Y_1, Y_2, $\ldots, Y_n$ is a random sample from the above distribution; we think of Y_i as independent identically distributed random variables.

- $y_1, \ldots, y_n$ are the actual observed values (the observed sample); these are non-random.

- $\widehat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ constructed from the sample.

Maximum likelihood for discrete distributions

Let $p(y;\boldsymbol{\theta})$ be a discrete probability distribution that depends on parameter $\boldsymbol{\theta}$. Given $\boldsymbol{\theta}$, the **likelihood function**, $\ell(\boldsymbol{\theta})$ equals the corresponding probability of the observed data:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n\,;\boldsymbol{\theta})\\ &= P(Y_1 = y_1\,;\boldsymbol{\theta})\,P(Y_2 = y_2\,;\boldsymbol{\theta})\,\ldots\,P(Y_n = y_n\,;\boldsymbol{\theta})\\ &= \prod_{i=1}^{n} p(y_i;\boldsymbol{\theta})\end{aligned}$$

Here $\boldsymbol{\theta}$ is the argument of the likelihood function and y_i are fixed.

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximises $\ell(\boldsymbol{\theta})$.

Maximum likelihood for continuous distributions

Let $p(y; \boldsymbol{\theta})$ be a density function. Given $\boldsymbol{\theta}$, the likelihood equals the corresponding density function, evaluated at the observed data:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= p(y_1, y_2, ..., y_n; \boldsymbol{\theta}) \\ &= p(y_1; \boldsymbol{\theta})\, p(y_2; \boldsymbol{\theta}) \ \dots \ p(y_n; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} p(y_i; \boldsymbol{\theta})\end{aligned}$$

Again, $\boldsymbol{\theta}$ is the argument of the likelihood function and y_i are fixed.

The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximises $\ell(\boldsymbol{\theta})$.

Log-likelihood

The log-likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log \ell(\boldsymbol{\theta}) \\ &= \log \left(\prod_{i=1}^{n} p(y_i; \boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^{n} \log p(y_i; \boldsymbol{\theta}) \end{aligned}$$

Because $L(\boldsymbol{\theta})$ is a monotonic transformation of $\ell(\boldsymbol{\theta})$, maximising the log-likelihood leads to the same solution, $\widehat{\boldsymbol{\theta}}$, as when maximising the likelihood.

Log-likelihood is often easier to work with than likelihood.

Example: Bernoulli distribution

Suppose that $Y_1, \ldots, Y_n$ come from the Bernoulli distribution with parameter θ (i.e. $Y_i = 1$ with probability θ and $Y_i = 0$ with prob. $1 - \theta$).

Note that we can write:

$$p(y_i; \theta) = P(Y_i = y_i) = \theta^{y_i}(1-\theta)^{(1-y_i)}$$

Thus,

$$\ell(\theta) = \prod_{i=1}^{n} p(y_i; \theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{(1-y_i)}$$

We now take the $\log$ to get the log-likelihood:

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{n} \left[y_i \log(\theta) + (1-y_i)\log(1-\theta) \right] \\ &= \left(\sum_{i=1}^{n} y_i \right) \log(\theta) + \left(n - \sum_{i=1}^{n} y_i \right) \log(1-\theta) \end{aligned}$$

Example: Bernoulli distribution

Derivative of the log-likelihood with respect to θ:

$$\frac{dL(\theta)}{d\theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1 - \theta}$$

Setting the derivative to zero, the MLE, $\widehat{\theta}$ must satisfy:

$$\frac{\sum_{i=1}^{n} y_i}{\widehat{\theta}} = \frac{n - \sum_{i=1}^{n} y_i}{1 - \widehat{\theta}}$$

The solution is the sample proportion:

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

Example: Gaussian MLR

We will treat the x values as fixed (i.e. non-random), and focus on the estimation of the $\boldsymbol{\beta}$. Recall that the Gaussian linear regression model specifies that Y_i are independent $N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\,,\, \sigma^2\right)$.

Using $\propto$ to denote "proportional to" and leaving out positive multiplicative constants we can write the likelihood as follows:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \prod_{i=1}^{n} p(y_i;\boldsymbol{\beta}) \\
&\propto \prod_{i=1}^{n} \exp\Big\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2\Big\} \\
&= \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2\Big\}.
\end{aligned}$$

Example: Gaussian MLR

$$\ell(\boldsymbol{\beta}) \;\propto\; \exp\Big\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\Big\}.$$

Maximizing the above expression over $\boldsymbol{\beta}$ is equivalent to maximizing the part in the exponent:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

which is equivalent to minimizing the residual sum of squares.

Thus, under the Gaussian multiple linear regression model, the MLE and the OLS estimators of the regression coefficients are identical.

Large sample properties of the ML estimator

- The MLE converges to the true parameter value as $n \to \infty$.
- The MLE is asymptotically unbiased (if there is a bias, it goes to zero as $n \to \infty$).
- The MLE is asymptotically optimal: it has the smallest variance (as $n \to \infty$) of any asymptotically unbiased estimator.

Bayesian approach

Bayesian inference

Recall that in the classical statistics the true parameter is fixed (nonrandom).

In Bayesian statistics, however, the parameter $\boldsymbol{\theta}$ can be treated as random, and we make inference about it conditional on the data.

Bayesian inference

In Bayesian inference, in addition to a sampling model $p(\boldsymbol{y}|\boldsymbol{\theta})$ we specify a **prior distribution** $p(\boldsymbol{\theta})$, which represents our beliefs about the parameter $\boldsymbol{\theta}$ before we see any data.

The Bayesian approach computes the **posterior distribution** $p(\boldsymbol{\theta}|\boldsymbol{y})$, which represents our updated beliefs about $\boldsymbol{\theta}$ after we observe the data $\boldsymbol{y}$.

As before, $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\boldsymbol{y})$ denote either probability mass functions or probability densities, depending on the context.

Posterior distribution

It follows from **Bayes' theorem** that the posterior distribution satisfies:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

Again using the $\propto$ notation and leaving out multiplicative constants that do not depend on $\boldsymbol{\theta}$ we can write the posterior as follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

We can restate the above relationship in words:

Posterior is proportional to Likelihood times Prior

Example: Gaussian MLR

We continue with the earlier example. Recall that the Gaussian linear regression model specifies that Y_i are independent $N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \, , \, \sigma^2\right)$.

We have shown that the likelihood has the following form:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) \;\propto\; \exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \right\}.$$

Example: prior

In the Bayesian approach, we need to choose a prior. We will assume that β_j are i.i.d. $N(0, \tau^2)$, for some $\tau^2 > 0$.

Note that this prior satisfies

$$p(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{j=0}^{p}\beta_j^2\right\}.$$

Example: posterior

Consequently,

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta})$$

$$\propto \exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\Big\}$$

$$\times \exp\Big\{ -\frac{1}{2\tau^2}\sum_{j=0}^{p}\beta_j^2\Big\}$$

$$\propto \exp\Big\{ -\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \frac{\sigma^2}{\tau^2}\sum_{j=0}^{p}\beta_j^2\Big]\Big\}$$

Example: posterior

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \quad \propto \quad \exp\Big\{-\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\frac{\sigma^2}{\tau^2}\sum_{j=0}^{p}\beta_j^2\Big]\Big\}.$$

We claimed last week that the mode of this posterior density corresponds to the Ridge regression estimator. We will show this shortly.

The function in the exponent is quadratic in $\boldsymbol{\beta}$. In fact, it can be shown that the posterior distribution of $\boldsymbol{\beta}$ is Gaussian (we will not worry about the proof of this fact).

The mode (and, thus, the mean) of this distribution converges to the OLS estimator $\widehat{\boldsymbol{\beta}}$ as the variance of the prior, τ^2 goes to infinity (i.e. the prior becomes less and less informative).

The mode converges to zero as the variance of the prior goes to zero (i.e. the prior becomes more and more concentrated around $\boldsymbol{\beta}=0$).

Maximum a posteriori (MAP) estimation

The **maximum a posteriori (MAP) estimator** is the mode of the posterior distribution:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{\text{MAP}} &= \underset{\boldsymbol{\theta}}{\text{argmax}}\; p(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \underset{\boldsymbol{\theta}}{\text{argmax}}\; \log(p(\boldsymbol{\theta}|\boldsymbol{y})) \\ &= \underset{\boldsymbol{\theta}}{\text{argmax}}\; \Big[\log(p(\boldsymbol{y}|\boldsymbol{\theta})) + \log(p(\boldsymbol{\theta}))\Big]\end{aligned}$$

Inside the square brackets we have the log-likelihood plus the log-prior. Incorporating the prior can be thought of as regularisation, and may reduce overfitting.

Many regularised risk minimisation methods have an interpretation as MAP estimation, without necessarily being fully Bayesian.

Example: Ridge estimator as a MAP estimator

Recall the posterior of Gaussian MLR

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \quad \propto \quad \exp\Big\{ -\frac{1}{2\sigma^2}\Big[\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\frac{\sigma^2}{\tau^2}\sum_{j=0}^{p}\beta_j^2\Big]\Big\}.$$

Let $\lambda = \sigma^2/\tau^2$, then the logarithm of the posterior density is:

$$\log\left[p(\boldsymbol{\beta}|\boldsymbol{y})\right] = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}\beta_j^2\right]+\text{constant},$$

where the "constant" comes from taking the $\log$ of the multiplicative factors that were left out in the expressions above.

The relevant part of the log-posterior consists of the ridge objective function times a negative multiplier. Hence, maximising the log-posterior is equivalent to *minimising* the ridge objective function. The MAP estimator is equivalent to the ridge estimator.

Introduction to classification

Classification

Consider the following business decisions.

1. Should we invest resources in acquiring and retaining a customer?
2. Should we offer a mortgage to a credit applicant?
3. Should we place a bid to sponsor an online search?
4. Should we investigate a transaction for possible fraud?

Classification

All these scenarios involve a **classification task**.

1. Do we predict that the customer will be profitable?
2. Do we predict that the applicant will repay the mortgage in full?
3. Do we predict that the user will click on the ad and make a purchase?
4. Do we flag the transaction?

Classification

In classification, the response variable Y is **qualitative** or **categorical** that takes values in a finite unordered set $\mathcal{Y} = \{1, \ldots, C\}$, where C is the number of classes. Our task is to predict which class a subject belongs to based on input variables.

A **classifier** $\widehat{y}$ is a mapping from the values of the inputs (predictors) to $\{1, \ldots, C\}$. A classifier is a prediction rule that assigns the subject to one of the classes, given the observed values of the predictors.

Given the value of the input vector (i.e. $X = \boldsymbol{x}$), we will write $\widehat{y}(\boldsymbol{x})$ (or simply $\widehat{y}$) for the value of the classifier.

Classification

In the fraud detection example, our response values may be: $\{\text{fraud}, \text{legitimate}\}$. The most common coding of binary variables is using the values 0 and 1, so we may code this variable as

$$Y = \begin{cases} 1 & \text{if fraud,} \\ 0 & \text{if legitimate.} \end{cases}$$

Notation

- Integers, such as 1,2,3 or 0,1, are used to denote the class labels.

- P, as in $P(A)$ or $P(Y = y)$, denotes a probability.

- p, as in $p(y)$ or $p(y|x)$, denotes a probability mass function (pmf) or probability density function (pdf), as before, depending on context.

Introduction to decision theory for classification

Loss functions (reminder)

A loss function $L(y, \widehat{y})$ measures the loss (or cost) of making a prediction $\widehat{y}$ when the truth is y. The most common loss function for regression is the squared error loss:

$$L(y, \widehat{y}) = \left(y - \widehat{y}\right)^2$$

For classification, the most popular loss function is the **0-1 loss**:

$$L(y, \widehat{y}) = \begin{cases} 1 \text{ if } y \neq \widehat{y} \\ 0 \text{ if } y = \widehat{y}. \end{cases}$$

The loss is zero for a correct classification and one for a misclassification.

Expected Loss

Given a classifier $\widehat{y}$, our objective is (as before) to minimise the corresponding expected loss:

$$E\left[L\Big(Y, \widehat{y}(X)\Big)\right]$$

the expectation is over Y and X, while the classifier $\widehat{y}$ is treated as fixed

You can think of the above quantity as the average loss across all subjects in the population (each subject has a Y and an X value).

Conditioning on the values of the predictors (i.e. on $X = \boldsymbol{x}$), we can write the expected loss as:

$$\sum_{c=1}^{C} L\Big(c, \widehat{y}(\boldsymbol{x})\Big) P(Y = c | X = \boldsymbol{x})$$

Bayes classifier

For the zero-one loss, the expected loss simplifies to:

$$
\begin{aligned}
\sum_{c=1}^{C} I\Big(c \neq \widehat{y}(\boldsymbol{x})\Big) P(Y = c|X = \boldsymbol{x}) \\
&= P\Big(Y \neq \widehat{y}(\boldsymbol{x})|X = \boldsymbol{x}\Big) \\
&= 1 - P\Big(Y = \widehat{y}(\boldsymbol{x})|X = \boldsymbol{x}\Big)
\end{aligned}
$$

Minimising this quantity is equivalent to choosing $\widehat{y}(\boldsymbol{x})$ that maximises the probability $P\Big(Y = \widehat{y}(\boldsymbol{x})|X = \boldsymbol{x}\Big)$

The solution is called the **Bayes classifier**, which classifies each subject to the most probable class.

Bayes error rate

Formally, the **Bayes classifier** is defined as:

$$\widehat{y}(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\text{argmax}}\ P(Y = c | X = \boldsymbol{x})$$

The **Bayes error rate** is the expected zero-one loss (i.e. the probability of misclassifying a test observation) for the Bayes classifier.

By definition, Bayes classifier has the lowest possible probability of misclassification. However, it requires knowing the distribution of Y given X.

Bayes decision boundary

Bayes decision boundary between two classes, say 0 and 1, is the set:

$$\{\, \boldsymbol{x} : \; P(Y = 0 | X = \boldsymbol{x}) = P(Y = 1 | X = \boldsymbol{x}) \,\}$$

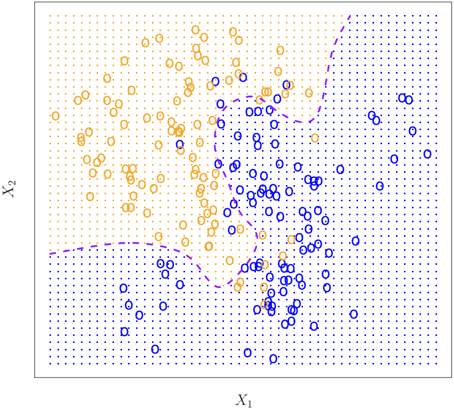


figure from ISL

Model Evaluation

Consider a classifier $\widehat{y}$ and a training dataset $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. The **training error rate** of this classifier is defined as:

$$\frac{1}{n}\sum_{i=1}^{n} I\Big(y_i \neq \widehat{y}(\boldsymbol{x}_i)\Big).$$

This gives the proportion of misclassifications on the training set.

When the above quantity is computed using a test set instead, it is called the **test error rate**. The Bayes classifier achieves the lowest expected test error rate (equivalently, the lowest test error rate over an infinitely large test set).

Classification

To approximate the Bayes classifier, we will use classification models and estimate conditional probabilities $\widehat{P}(Y = c|X = \boldsymbol{x})$ for $c = 1, \ldots, C$. We will then classify a subject to the class with highest estimated probability.

In particular, in binary classification with the 0 - 1 coding for Y, we make a prediction $\widehat{y}(\boldsymbol{x}) = 1$ if $\widehat{P}(Y = 1|X = \boldsymbol{x}) > 0.5$.

Otherwise, we make a prediction $\widehat{y}(\boldsymbol{x}) = 0$.

K-nearest neighbours classifier

K-nearest neighbours classifier

Given training data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ and an input point $\boldsymbol{x}$,

K-nearest neighbours classifier estimates the conditional probability for class c as:

$$\widehat{P}(Y = c|X = \boldsymbol{x}) = \text{Average}\Big[\, I(y_i = c) \,\Big|\, \boldsymbol{x}_i \text{ is in } \mathcal{N}_k(x) \,\Big]$$

Here we average $I(y_i = c)$ for the observations whose $\boldsymbol{x}_i$ lie in the neighborhood $\mathcal{N}_k(\boldsymbol{x})$ containing the closest k data points to $\boldsymbol{x}$.

K-nearest neighbours classifier

Given training data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$ and an input point $\boldsymbol{x}$,

K-nearest neighbours classifier estimates the conditional probability for class c as:

$$\widehat{P}(Y = c|X = \boldsymbol{x}) = \text{Average}\Big[\, I(y_i = c) \,\Big|\, \boldsymbol{x}_i \text{ is in } \mathcal{N}_k(x) \,\Big]$$

Here we average $I(y_i = c)$ for the observations whose $\boldsymbol{x}_i$ lie in the neighborhood $\mathcal{N}_k(\boldsymbol{x})$ containing the closest k data points to $\boldsymbol{x}$.

Thus, KNN finds the K training input points that are closest to $\boldsymbol{x}$ and then estimates $P(Y = c|X = \boldsymbol{x})$ as the proportion (i.e. fraction) of these K points that belongs to the class c.

Illustration: KNN with $K = 3$

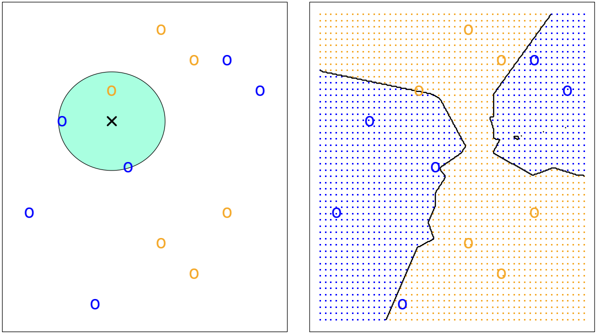


figure from ISL

K-nearest neighbours classifier

- KNN classifier is a direct nonparametric approximation to the Bayes classifier.

- The lower the K, the more flexible the decision boundary.

- As always, choosing the optimal level of flexibility is crucial. We use cross validation to select K.

Example: KNN vs Bayes decision boundaries

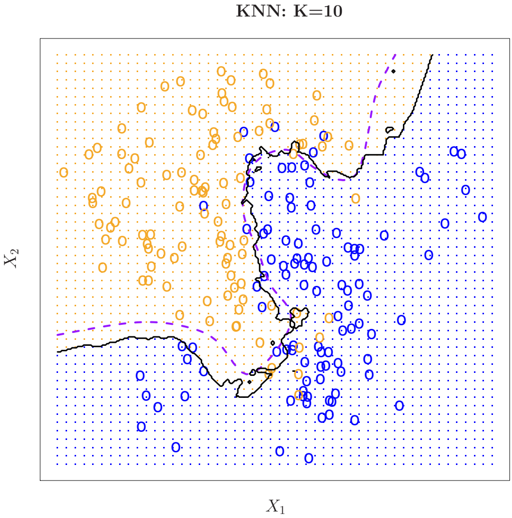


figure from ISL

Example: KNN vs Bayes decision boundaries

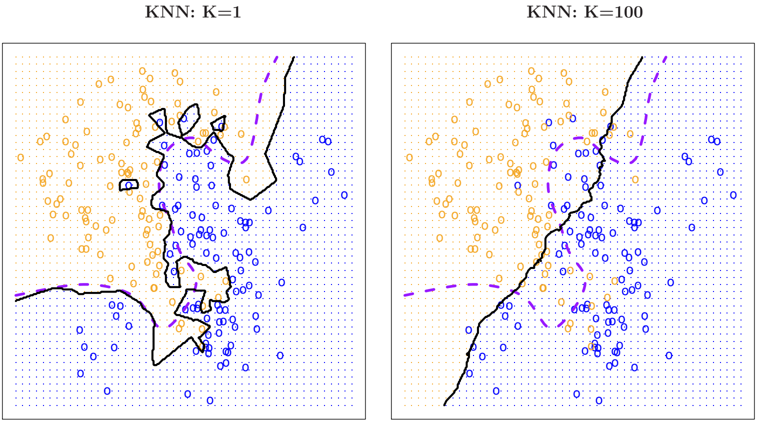
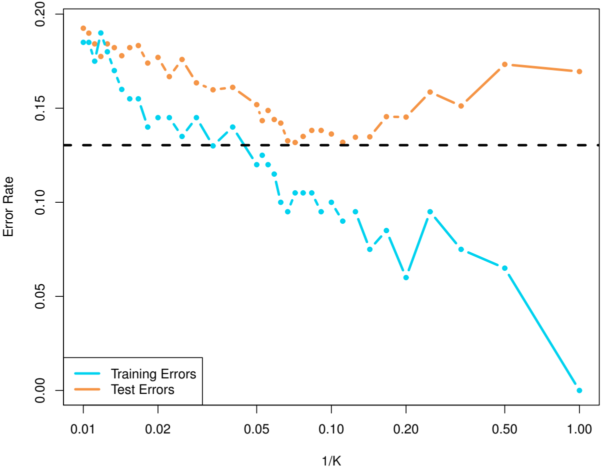


figure from ISL

Example: KNN error rates



Black dashed line gives the Bayes error rate.

Review questions

- What is regularised risk minimisation?
- What is maximum likelihood estimation?
- What is Bayesian statistics?
- What is a posterior distribution?
- What is a zero-one loss?
- What is the Bayes classifier?
- What is the test error rate?
- Explain the KNN classifier.