QBUS6810: Statistical Learning and Data Mining

Lectures 7-8: Classification I

Semester 1, 2019

Discipline of Business Analytics, The University of Sydney Business School

Lectures 7-8: Classification I

1. Classification
2. Introduction to decision theory for classification
3. K-nearest neighbours classifier
4. Review of the Bayes' rule
5. Naïve Bayes classifier
6. Decision theory for binary classification

Classification

Classification

Consider the following business decision making scenarios.

1. Should we invest resources in acquiring and retaining a customer?
2. Should we offer a mortgage to a credit applicant?
3. 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 flag the transaction?

Classification

In classification, the response variable Y is **qualitative** or **categorical** that takes values in a finite unordered set.

In general, we can think of Y as a variable taking values in the set $\mathcal{Y} = \{1, \ldots, C\}$, where C is the number of the corresponding classes. Our task is to predict which class a subject belongs to based on the 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, the response values are: $\{\text{fraudulent}, \text{legitimate}\}$. The most common coding for such binary variables is using the values 0 and 1:

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

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).

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**:

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

Here $I(\cdot)$ is the indicator function. The zero-one 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 nonrandom

We 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).

Expected Loss

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

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

Here $\boldsymbol{x}$ and $\widehat{y}(\boldsymbol{x})$ are fixed.

In the case of the zero-one loss, the above expression becomes

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

In other words, we are summing the probabilities $P(Y = y | X = \boldsymbol{x})$ over all values y that are different from $\widehat{y}(\boldsymbol{x})$.

Bayes classifier

Thus, the conditional expected zero-one loss is given by:

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

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 corresponding solution is called the **Bayes classifier**, which classifies each subject to the most probable (most likely) class.

Bayes error rate

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

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

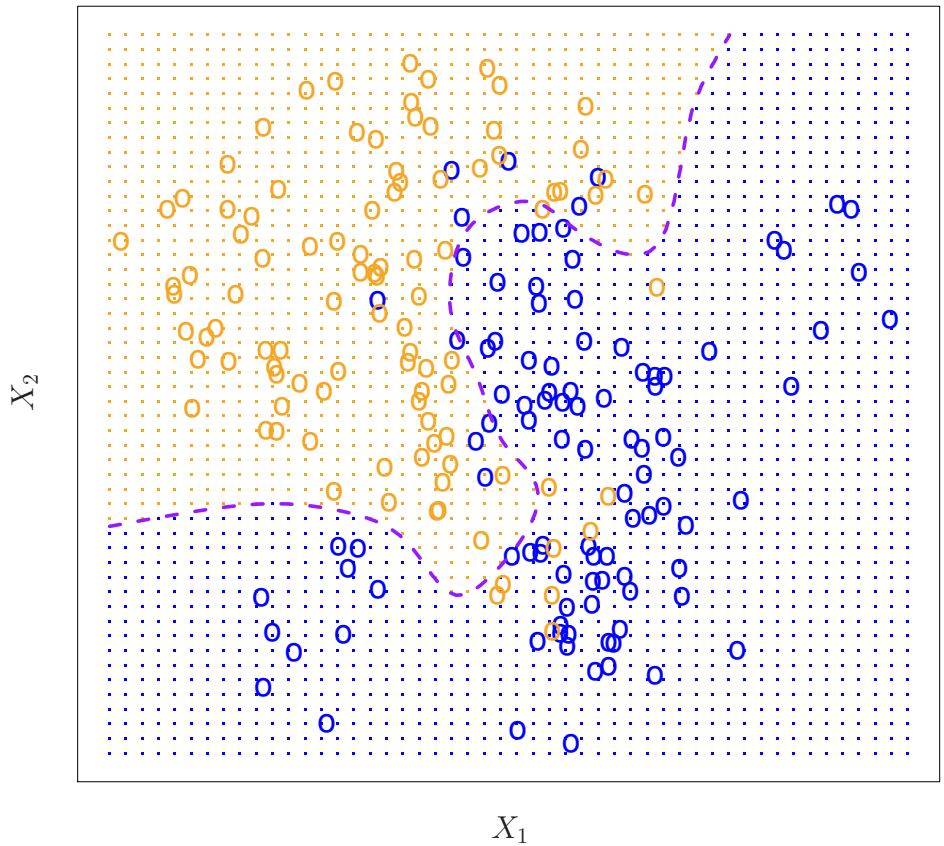
The **Bayes error rate** is the expected zero-one loss (i.e. the probability of misclassifying a test observation) for the Bayes classifier.

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:

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



The blue class and the orange class are equally likely for x on the Bayes decision boundary. The probability of the blue class is greater than 0.5 for x in the blue region and lower than 0.5 in the orange region.

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 (informally: 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 $\widehat{P}(Y = y|X = \boldsymbol{x})$ for $y = 1, \ldots, C$. We will then classify a subject to the class with the 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 y as:

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

Here we average $I(y_i = y)$ 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 $P(Y = y|X = \boldsymbol{x})$ as the proportion (i.e. fraction) of these K points that belongs to the class y.

Illustration: KNN with $K = 3$

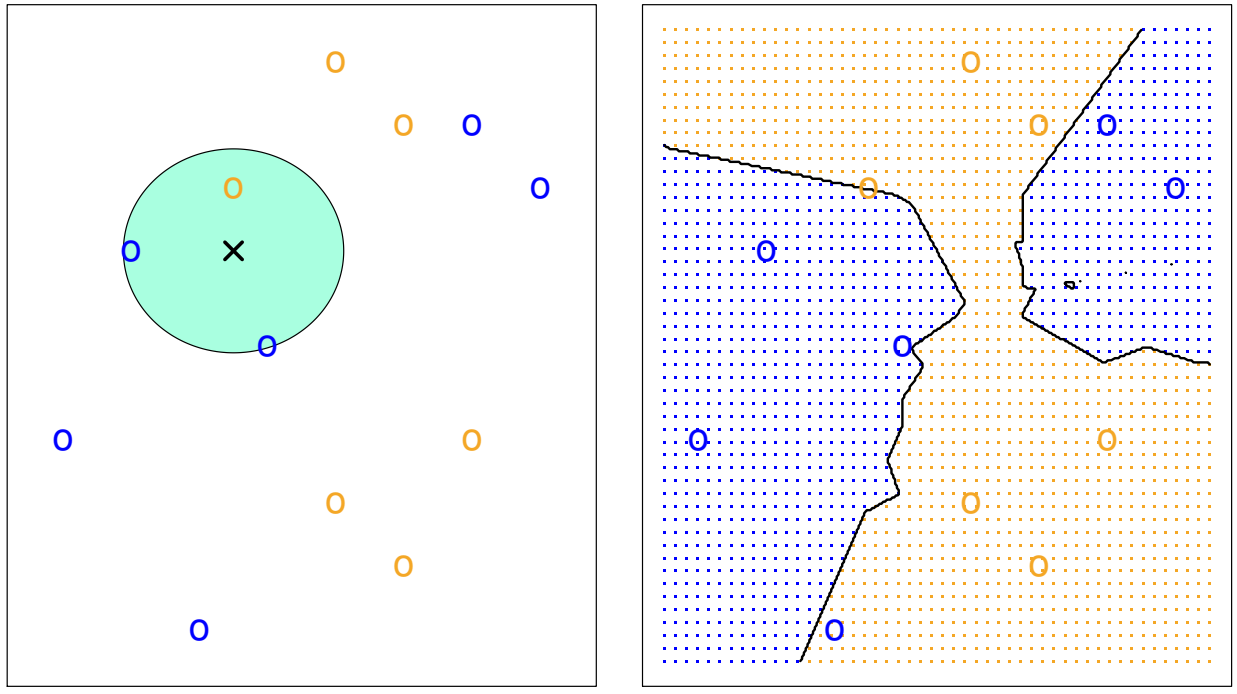


figure from ISL

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

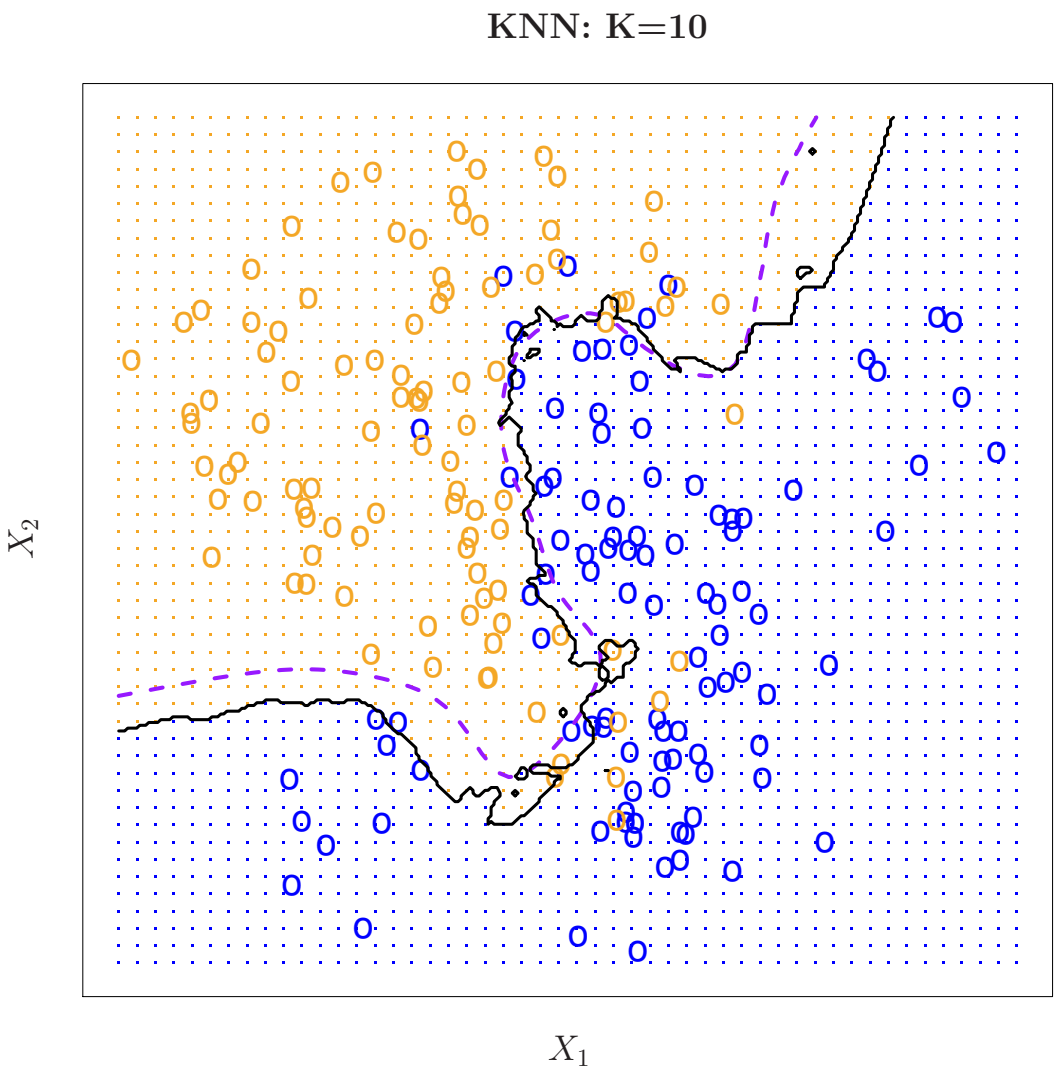
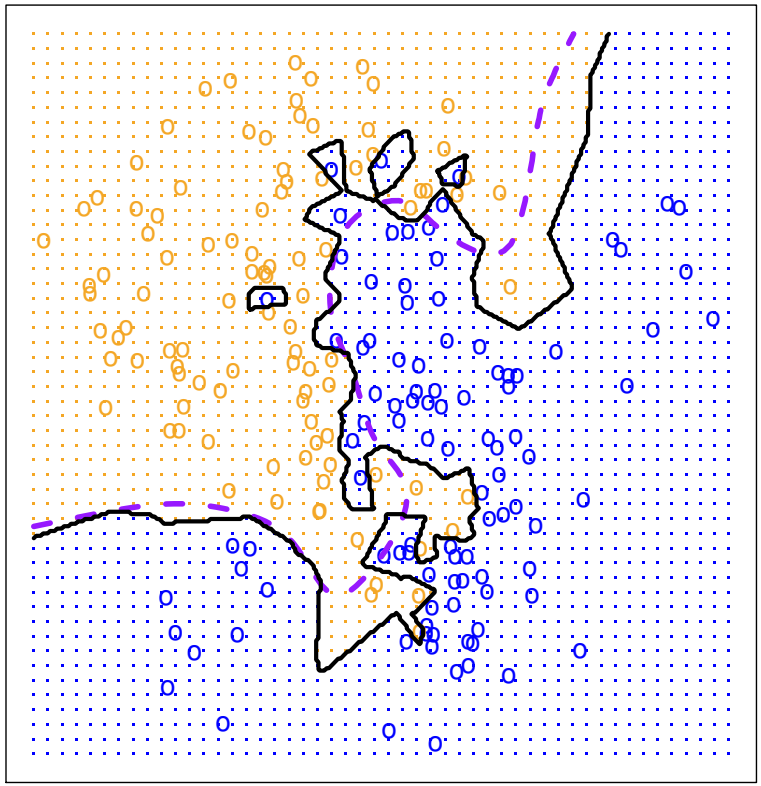


figure from ISL

Example: KNN vs Bayes decision boundaries

KNN: K=1



KNN: K=100

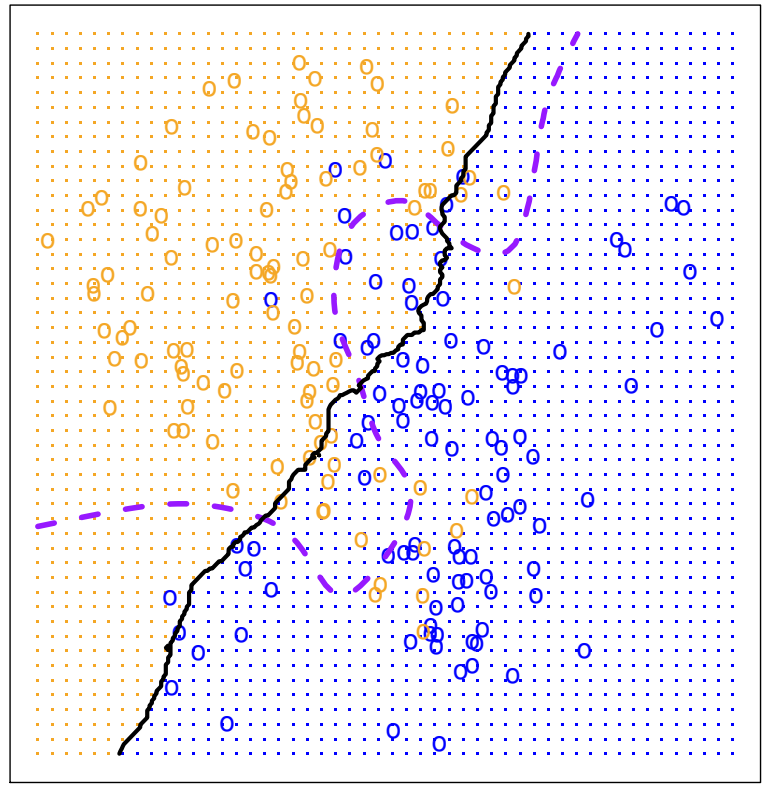
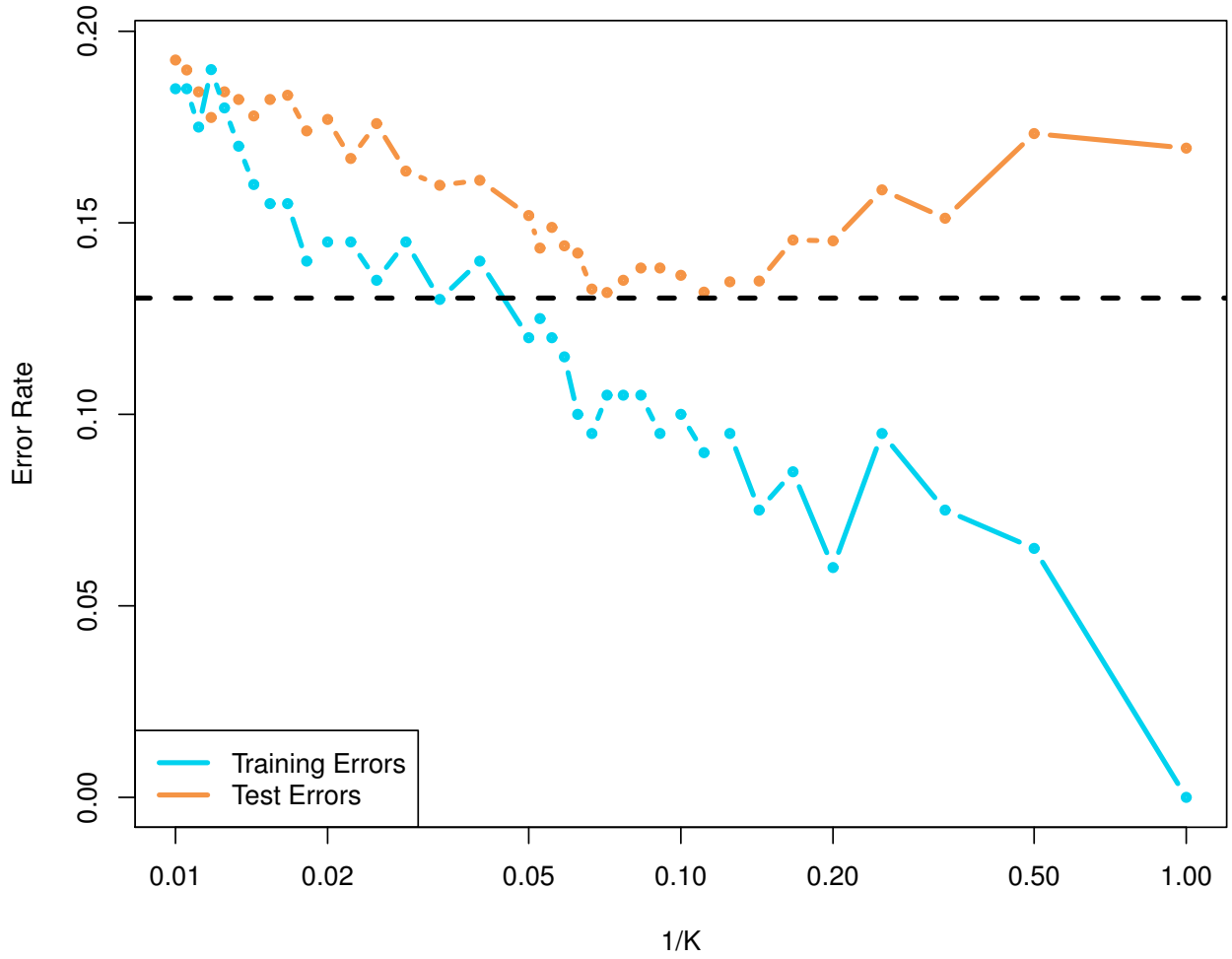


figure from ISL

Example: KNN error rates



Black dashed line gives the Bayes error rate.

Our next topic is the Naïve Bayes classifier.

Before we discuss this method, we will review a useful probability rule, called the Bayes' rule.

Review of the Bayes' rule

Notation

Suppose that Y takes values in the set $\{1, 2, \ldots, C\}$.

Define $\pi_y = P(Y = y)$.

If X is continuous, write $p(\boldsymbol{x}|Y = y)$ for the density of X conditional on $Y = y$.

If X is discrete, let $p(\boldsymbol{x}|Y = y)$ denote the conditional probability $P(X = x|Y = y)$.

Bayes' rule

Let X be discrete and recall that we defined $\pi_y = P(Y = y)$.

The **Bayes' rule** or **Bayes' theorem** gives:

$$P(Y = y|X = \boldsymbol{x}) = \frac{P(X = \boldsymbol{x}|Y = y)\pi_y}{P(X = \boldsymbol{x}|Y = 1)\pi_1 + P(X = \boldsymbol{x}|Y = 2)\pi_2 + \ldots + P(X = \boldsymbol{x}|Y = C)\pi_C}$$

Using our notation, we can write the last expression as:

$$\frac{p(\boldsymbol{x}|Y = y)\pi_y}{p(\boldsymbol{x}|Y = 1)\pi_1 + p(\boldsymbol{x}|Y = 2)\pi_2 + \ldots + p(\boldsymbol{x}|Y = C)\pi_C}$$

Bayes' rule

Now let X be continuous and recall that we write $p(\boldsymbol{x}|Y = y)$ for the density of X conditional on $Y = y$.

Similarly to the discrete case, the Bayes' rule gives:

$$P(Y = y|X = \boldsymbol{x})$$

$$= \frac{p(\boldsymbol{x}|Y = y)\pi_y}{p(\boldsymbol{x}|Y = 1)\pi_1 + p(\boldsymbol{x}|Y = 2)\pi_2 + \ldots + p(\boldsymbol{x}|Y = C)\pi_C}$$

Note that the last expression is identical to the one in the discrete case (previous slide)

Example: medical test

Consider a medical test for cancer. Suppose that the test has a sensitivity of 80%, which means that if a person has cancer, the test will return positive with probability 0.8:

$$P(X = 1|Y = 1) = 0.8.$$

Here X is the outcome of the test and Y is the indicator of the presence of cancer.

In case of a positive test result, what is the probability that a person has cancer, i.e. what is $P(Y = 1|X = 1)$?

Example: medical test

Using Bayes' theorem:

$$P(Y = 1|X = 1)$$

$$= \frac{P(X = 1|Y = 1)\pi_1}{P(X = 1|Y = 1)\pi_1 + P(X = 1|Y = 0)\pi_0}$$

This equation tells us that in order to calculate the desired probability, we also need to know the probability of cancer (π_1) and the so-called false positive rate: $P(X = 1|Y = 0)$.

Example: medical test

Suppose that $\pi_1 = 0.004$ and $P(X = 1|Y = 0) = 0.1$.

Also, recall that we assumed $P(X = 1|Y = 1) = 0.8$. Then:

$$P(Y = 1|X = 1)$$

$$= \frac{P(X = 1|Y = 1)\pi_1}{P(X = 1|Y = 1)\pi_1 + P(X = 1|Y = 0)\pi_1}$$

$$= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Naïve Bayes classifier

Naïve Bayes classifier

In the classification setting, we refer to $p(\boldsymbol{x}|Y = y)$ as the class conditional densities (or probability mass functions if X is discrete), and we refer to $\pi_y = P(Y = y)$ as class probabilities.

The Naïve Bayes classifier method uses the general approach of modeling the conditional distribution of X given Y, and then using the Bayes' rule to obtain the conditional distribution of Y given X:

$$P(Y = y|X = \boldsymbol{x})$$

$$= \frac{p(\boldsymbol{x}|Y = y)\pi_y}{p(\boldsymbol{x}|Y = 1)\pi_1 + p(\boldsymbol{x}|Y = 2)\pi_2 + \ldots + p(\boldsymbol{x}|Y = C)\pi_C}$$

Naïve Bayes classifier

The **Naïve Bayes classifier** (NBC) is based on the assumption that the predictors are conditionally independent given the class label.

Thus, the class conditional density (or probability mass function if X is discrete) factorises into a product of the individual predictor densities:

$$p(\boldsymbol{x}|Y = y) = \prod_{j=1}^{p} p(x_j|Y = y).$$

Naïve Bayes classifier

- The method is "naive" because we do not think that the features are in fact conditionally independent.
- The simplicity of the NBC method makes it relatively immune to overfitting, which is useful for applications where the number of features is large.
- The assumption of conditional independence makes it easy to mix and match different predictor types.

Naïve Bayes classifier

- Despite being based on an assumption that is not necessarily true, the Naïve Bayes classifier often performs very well in practice compared to more complex alternatives.

- The reason is again the bias-variance trade-off: while the assumption of class-conditional independence may lead to biased probabilities, the simplifications brought by it may lead to substantial reduction in variance.

Continuous predictors

For real-valued predictors, a common assumption is that:

$$X_j|Y = y \quad \sim \quad N(\mu_{jy}, \sigma^2_{jy})$$

where μ_{jy} and σ^2_{jy} are the mean and the variance of predictor j conditional on the class y.

Parameters μ_{jy} and σ^2_{jy} need to be estimated from the data.

Continuous predictors

- Typically, we first transform the predictors in order to make the variables approximately normal or symmetric.

- We could also use other distributional assumptions or follow a nonparametric approach to estimate the class conditional densities.

Binary predictors

When the predictors are binary, i.e. X_j only takes values 0 and 1, we use the Bernoulli distribution:

$$X_j|Y = y \quad \sim \quad \text{Bernoulli}(\theta_{jy})$$

where θ_{jy} is the probability that $X_j = 1$ given $Y = y$

(and, thus, $1 - \theta_{jy}$ is the probability that $X_j = 0$ given $Y = y$)

Parameters θ_{jy} need to be estimated from the data.

Application: document classification

Document classification is the problem of classifying text documents into different categories.

A simple approach is to represent each document as a vector of binary variables, where each variable records whether a particular word is present in the document or not. For example, $x_{ij} = 1$ if the word j appears in document i, and $x_{ij} = 0$ otherwise.

This is called a **bag of words** model.

Estimating Naïve Bayes parameters using maximum likelihood

We estimate the parameters in the Naïve Bayes model by maximum likelihood. In particular, we:

1. Estimate the prior class probabilities π_y by computing the sample proportions of each class in the training data.

2. Fit univariate models and estimate the parameters separately for each predictor within each class (the fact that we can do this is a direct consequence of the assumption of conditional independence).

the next two slides give some mathematical details, but you will **not** need to reproduce them on the exam

Estimating Naïve Bayes parameters using maximum likelihood

Let $\boldsymbol{\theta}$ contain all the parameters for the class conditional densities of the predictors, and let $\boldsymbol{\pi}$ contain the class probabilities for Y.

The density (or probability) for observation i is

$$\begin{aligned} p(\boldsymbol{x}_i, y_i; \boldsymbol{\theta}, \boldsymbol{\pi}) &= p(\boldsymbol{x}_i | y_i; \boldsymbol{\theta}) p(y_i; \boldsymbol{\pi}) \\ &= \prod_{j=1}^{p} p(x_{ij} | y_i; \boldsymbol{\theta}_j) p(y_i; \boldsymbol{\pi}) \end{aligned}$$

Thus, the likelihood is:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{j=1}^{p} p(x_{ij} | y_i; \boldsymbol{\theta}_j) p(y_i; \boldsymbol{\pi}).$$

Estimating Naïve Bayes parameters: class probabilities

The likelihood is:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \prod_{j=1}^{p} p(x_{ij}|y_i; \boldsymbol{\theta}_j) p(y_i; \boldsymbol{\pi}).$$

Hence, the log-likelihood is:

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{j=1}^{p} \sum_{i=1}^{n} \log\left(p(x_{ij}|y_i; \boldsymbol{\theta}_j)\right) + \sum_{i=1}^{n} \log\left(p(y_i; \boldsymbol{\pi})\right)$$

Note that the log-likelihood decomposes into a sum of terms, each involving a different parameter: $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_p$, and $\boldsymbol{\pi}$. We can therefore maximize all these terms separately.

Estimating Naïve Bayes parameters using maximum likelihood

As a consequence:

$$\widehat{\pi}_y = \frac{n_y}{n}$$

where n is the sample size, as always, and $n_y = \sum_{i=1}^{n} I(y_i = y)$ is the number of observations that fall in class y.

Estimating Naïve Bayes parameters: binary predictors

Suppose that the predictors are binary, such that

$$X_j|Y = y \quad \sim \quad \text{Bernoulli}(\theta_{jy})$$

The MLE of each parameter $\theta_{jy} = P(X_j = 1|Y = y)$ is:

$$\widehat{\theta}_{jy} = \frac{n_{jy}}{n_y}$$

where $n_{jy} = \sum_{i=1}^{n} I(x_{ij} = 1)I(y_i = y)$ is the number of observations that fall in class y for which predictor j equals 1.

Estimating Naïve Bayes parameters: Gaussian case

Suppose that the class conditional distribution is Gaussian:

$$X_j|Y = y \quad \sim \quad N(\mu_{jy}, \sigma^2_{jy})$$

The MLEs are:

$$\widehat{\mu}_{jy} \quad = \quad \frac{1}{n_y} \sum_{i:\, y_i = y} x_{ij}$$

i.e. the sample mean of predictor X_j using just the subjects from class y

$$\widehat{\sigma}^2_{jy} \quad = \quad \frac{1}{n_y} \sum_{i:\, y_i = y} (x_{ij} - \widehat{\mu}_{jy})^2$$

i.e. the "sample variance" of X_j using just the subjects from class y

Decision theory for binary classification

Classification outcomes

In most business problems, there are distinct losses associated with each classification outcome. Consider, for example, the case of transaction fraud detection.

		Classification	
		Legitimate	**Fraud**
Actual	**Legitimate**	No loss	Investigation cost
	Fraud	Fraud loss	Fraud loss avoided

The cost of investigating a suspicious transaction is likely to be much lower than the loss in case of fraud.

Classification outcomes

We will use the following terminology.

		Classification	
		$\widehat{y} = 0$	$\widehat{y} = 1$
Actual	$Y = 0$	True negative	False positive
	$Y = 1$	False negative	True positive

Loss matrix

The context of the business problem will often specify a **loss matrix** or **cost-benefit matrix** for classification as follows.

		Classification	
		$\widehat{y} = 0$	$\widehat{y} = 1$
Actual	$Y = 0$	L_{TN}	L_{FP}
	$Y = 1$	L_{FN}	L_{TP}

Example: credit scoring

In credit scoring, we want to classify a loan applicant as creditworthy ($Y = 1$) or not ($Y = 0$) based on the probability that the customer will not default.

		Classification	
		$\widehat{y} = 0$	$\widehat{y} = 1$
Actual	$Y = 0$	Default loss avoided	Default loss
	$Y = 1$	Profit opportunity lost	Profit

A false positive is a more costly error than a false negative in this business scenario. Our decision making should therefore take this into account.

Decision rule

The decision to classify a subject as positive or negative is based on the following decision rule:

$$\widehat{y}(\boldsymbol{x}) = \begin{cases} 1 \text{ if } \widehat{P}(Y = 1 | X = \boldsymbol{x}) > \tau. \\ 0 \text{ if } \widehat{P}(Y = 1 | X = \boldsymbol{x}) \leq \tau. \end{cases}$$

Here $\widehat{P}$ corresponds to the estimated conditional probability, and τ is a decision threshold parameter. Recall that in the case of binary classification with the zero-one loss we use $\tau = 0.5$.

Optimal decision

It has been shown (but we will not go into the proof) that the optimal value of the threshold (the one minimising expected loss) is:

$$\tau^* = \frac{L_{\mathsf{FP}} - L_{\mathsf{TN}}}{L_{\mathsf{FP}} + L_{\mathsf{FN}} - L_{\mathsf{TP}} - L_{\mathsf{TN}}}$$

Example: zero-one loss

With the zero-one loss, we have that $L_{\mathsf{FP}} = L_{\mathsf{FN}} = 1$ and $L_{\mathsf{TP}} = L_{\mathsf{TN}} = 0$, i.e. the loss matrix is:

		Classification	
		$\widehat{y} = 0$	$\widehat{y} = 1$
Actual	$Y = 0$	0	1
	$Y = 1$	1	0

Therefore,

$$\tau^* = \frac{L_{\mathsf{FP}} - L_{\mathsf{TN}}}{L_{\mathsf{FP}} + L_{\mathsf{FN}} - L_{\mathsf{TP}} - L_{\mathsf{TN}}} = \frac{1}{2}$$ as before

Example: credit scoring

In the credit scoring example, we can set the loss matrix as:

		Classification	
		$\widehat{y}=0$	$\widehat{y}=1$
Actual	$Y=0$	0	L_{FP}
	$Y=1$	L_{FN}	0

where L_{FN} equals missed profit and L_{FP} equals default loss.

Therefore,

$$\tau^* = \frac{L_{\mathsf{FP}} - L_{\mathsf{TN}}}{L_{\mathsf{FP}} + L_{\mathsf{FN}} - L_{\mathsf{TP}} - L_{\mathsf{TN}}} = \frac{L_{\mathsf{FP}}}{L_{\mathsf{FP}} + L_{\mathsf{FN}}}.$$

Example: credit scoring

We've shown that the optimal threshold for the loan decision is:

$$\tau^* = \frac{L_{\mathsf{FP}}}{L_{\mathsf{FN}} + L_{\mathsf{FP}}}.$$

We expect the loss from default to be much higher than the profit from a loan to a creditworthy customer (L_{FP} much larger than L_{FN}). Note that this leads to a high value of the threshold τ^*.

In other words, it is only worth it to lend to customers that have a high probability of repayment.

Review questions

- What is a zero-one loss?
- What is the Bayes classifier?
- What is the test error rate?
- Explain the KNN classifier.
- What is the key assumption of the Naive Bayes classifier?
- What is a loss matrix, and what are its entries in the case of the zero-one loss?