Statistical Learning and Data Mining

Module 10: Classification I

Semester 2, 2017

Discipline of Business Analytics, The University of Sydney Business School

Module 10: Classification I

1. Classification
2. Review of probability theory
3. Introduction to decision theory for classification
4. K-nearest neighbours classifier
5. Naïve Bayes classifier
6. Decision theory for binary classification
7. Model evaluation 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 invest more resources to train an employee?
4. Should we place a bid to a sponsor an online search?
5. 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 employee will stay in the company?
4. Do we predict that the user will click on the ad and make a purchase?
5. 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** $\widehat{Y}(X)$ is a mapping from the input vector $\boldsymbol{x}$ 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.

Classification

In the fraud detection example, our response variable may be flag $= \{\text{fraud}, \text{legitimate}\}$. We can code this variable as

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

Notation

- We use numbers as a generic notation for the class labels.
- A capital P as in $P(A)$ or $P(Y = y)$ denotes a probability.
- A lower case p as in $p(y)$ or $p(y|x)$ denotes a probability mass function (pmf) or probability density function (pdf).

Review of probability theory

Discrete random variables

A **discrete random variable** X is a random variable that takes a value in a finite or countably infinite set $\mathcal{X}$, for example $\mathcal{X} = \{0, 1\}$, $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, or $\mathcal{X} = \mathbb{N}$.

We denote the probability of the event that $X = x$ by $P(X = x)$ or $p(x)$ for short. We call $p(x)$ the **probability mass function** or **pmf**, where $0 \leq p(x) \leq= 1$ and the probabilities sum to one over $\mathcal{X}$.

Probabilities of unions of events and joint events

Given two events A and B, we define the probability of A or B as

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

which equals to $p(A) + p(B)$ if A and B are mutually exclusive.

We define the probability of the joint event A and B as

$$P(A, B) = P(A \cap B) = P(A|B)P(B),$$

which we refer to as the **product rule**.

Marginal distribution

Given the joint distribution $p(x, y)$, we define the **marginal distribution** of X as

$$\begin{aligned} P(X = x) &= \sum_{y \in \mathcal{Y}} P(X = x, Y = y) \\ &= \sum_{y \in \mathcal{Y}} P(X = x | Y = y) P(Y = y), \end{aligned}$$

where we sum over all the possible states of Y.

We call this definition the **sum rule**.

Conditional probability

We define the conditional probability of event A, given that event B is true, as

$$P(A|B) = \frac{P(A,B)}{P(B)},$$

if $P(B) > 0$.

Bayes' rule (key concept)

Let X and Y be two discrete random variables. We state the **Bayes' rule** or **Bayes' theorem** as

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

$$= \frac{P(X = x|Y = y)P(Y = y)}{\sum_{y' \in \mathcal{Y}} P(X = x|Y = y')P(Y = y')}$$

Example: medical test

A mammogram is a medical test for breast cancer. Suppose that the test has a **sensitivity** of 80%, which means that if a woman has cancer, the test will return positive with probability 0.8,

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

In case of a positive test result, what is the probability that a woman has cancer?

Example: medical test

Using Bayes' theorem, the conditional probability $P(Y = 1|X = 1)$ is

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

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

Ignoring the prevalence is a logical mistake known as the **base rate fallacy**.

Example: medical test

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

$$P(Y = 1|X = 1) = \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} = 0.031$$

Multinoulli distribution (key concept)

The **multinoulli** or **categorical distribution** is a generalisation of the Bernoulli distribution to multiple outcomes. A multinoulli random variable takes a value $y \in \{1, \ldots, C\}$.

Each outcome has probability

$$P(Y = c) = \pi_c$$

for $c = 1, \ldots, C$. The probabilities need to satisfy $\sum_{c=1}^{C} \pi_c = 1$.

Multinoulli distribution

We denote a multinoulli random variable as $Y \sim \text{Cat}(\pi_1, \ldots, \pi_C)$. The probability mass function is

$$p(y) = \prod_{c=1}^{C} \pi_c^{I(y=c)},$$

where $I(\cdot)$ is the indicator function. That means that $p(y) = \pi_c$ when $y = c$.

The vector of categorical counts from N independent trials of the multinoulli distribution follows the **multinomial distribution**.

Introduction to decision theory for classification

Loss functions for classification

In classification, we represent the loss function by a $C \times C$ loss matrix $\mathbf{L}$. Each element of the loss matrix $L_{k\ell} = L(k, \ell)$ specifies the loss of classifying in class ℓ when the actual class is k.

In this section, we focus on a simple framework by considering the **zero-one loss function**.

Zero-one loss function (key concept)

The zero-one loss function is

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

such that the loss is zero for a correct classification and one for a misclassification.

Classification risk

As before, our objective is to minimise the risk or expected loss of the classifier,

$$R(\widehat{Y}(X)) = E\left[L(Y, \widehat{Y}(X))\right].$$

You can think of the risk as the average loss across all subjects in the population (each subject has a pair of Y and X values).

By conditioning on the predictors, we can rewrite the risk as

$$R(\widehat{Y}(X)) = E_X\left(\sum_{c=1}^{C} E\left[L(c, \widehat{Y}(X))\right] P(Y = c|X)\right)$$

Bayes classifier (key concept)

It is sufficient to minimise the loss for any given input,

$$\widehat{y} = \underset{k \in \mathcal{Y}}{\text{argmin}} \sum_{c=1}^{C} L(c, k) P(Y = c | X = \boldsymbol{x}).$$

For the zero-one loss, this simplifies to

$$\widehat{y} = \underset{c \in \mathcal{Y}}{\text{argmin}} \ [1 - P(Y = c | X = \boldsymbol{x})]$$

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

Bayes error rate (key concept)

Formally, the Bayes classifier is

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

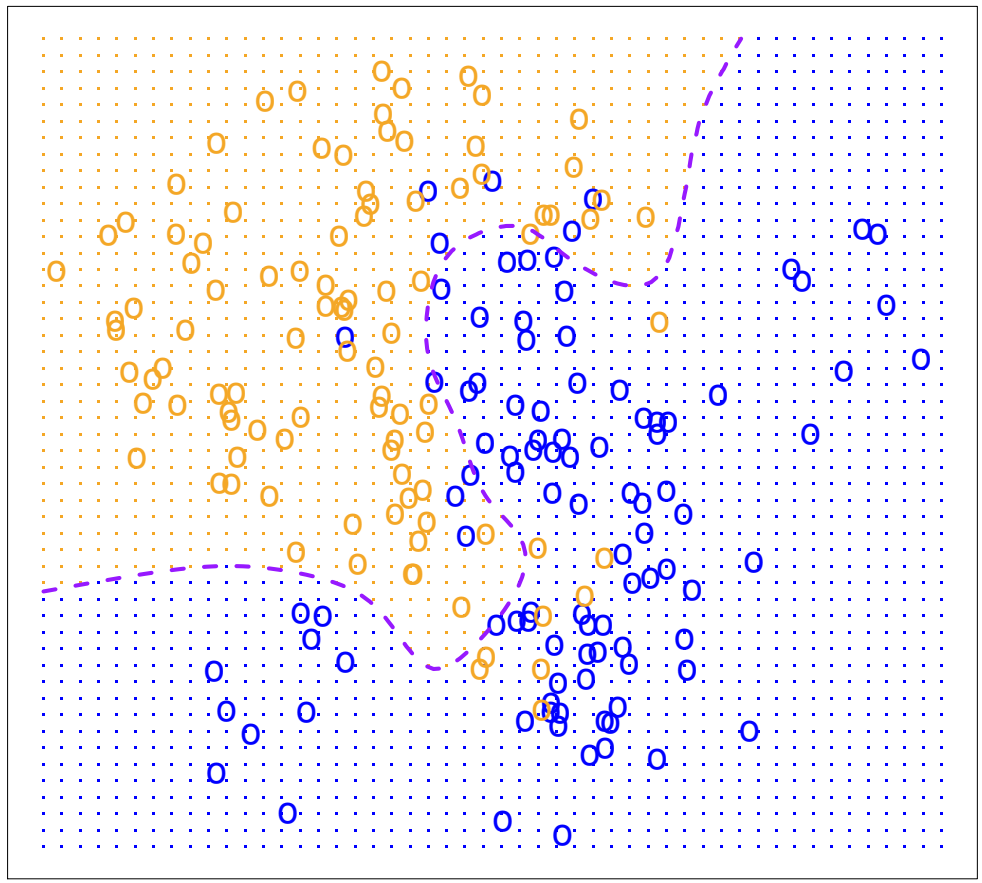
such that $P(Y = \widehat{y} | X = \boldsymbol{x}) \geq P(Y = c | X = \boldsymbol{x})$ for all $c \in \mathcal{Y}$.

The **Bayes error rate** is the expected zero-one loss under the Bayes classifier.

Bayes decision boundary

The **Bayes decision boundary** between classes k and ℓ is the set

$$\{\boldsymbol{x}:\ P(Y=k|X=\boldsymbol{x})=P(Y=\ell|X=\boldsymbol{x})\}\,.$$



Classification

Classification models lead to estimated conditional probabilities $\widehat{P}(Y = c|X = \boldsymbol{x})$ for $c = 1, \ldots, C$. We then classify a test case to the class with highest estimated probability.

For binary classification, we therefore classify a subject as $\widehat{y} = 1$ if $\widehat{P}(Y = 1|X = \boldsymbol{x}) \geq 0.5$.

Model evaluation

We compute the **misclassification** or **error rate** for the test data as

$$\overline{\mathsf{Err}}_{\mathsf{test}} = \frac{1}{n}\sum_{i=1}^{n} I(y_{i,0} \neq \widehat{y}_{i,0}).$$

K-nearest neighbours classifier

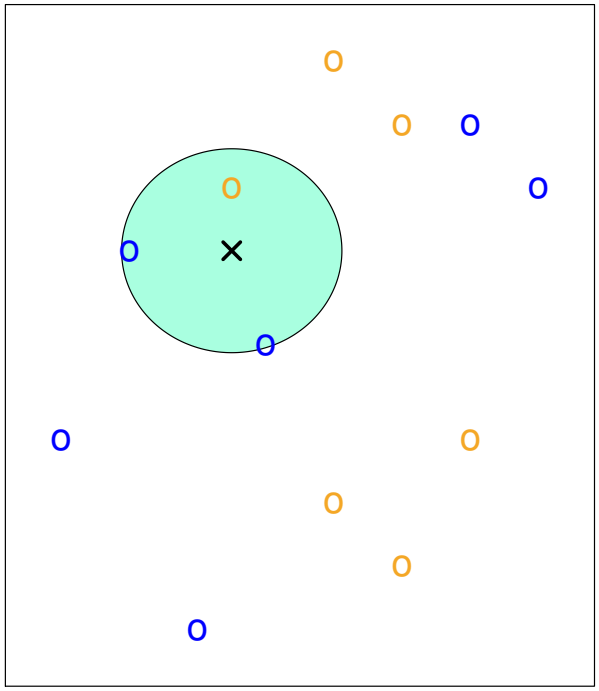
K-nearest neighbours classifier

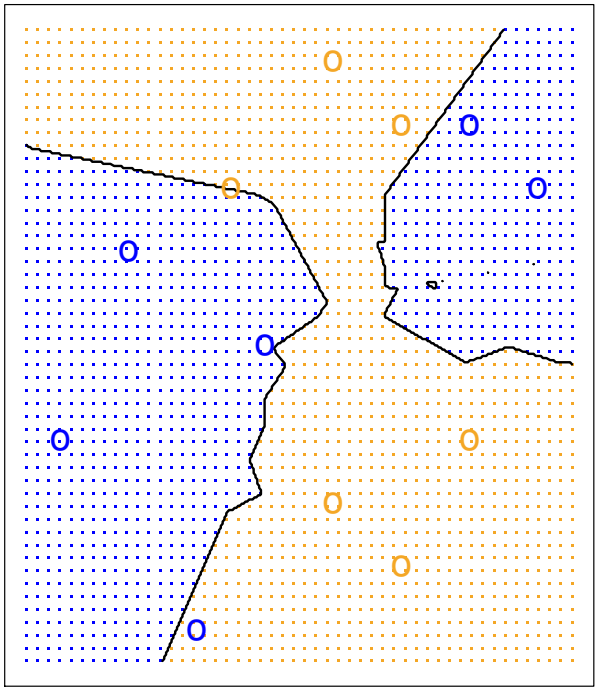
The **K-nearest neighbours classifier** estimates the conditional probability for class c as

$$\widehat{P}(Y = c|X = \boldsymbol{x}) = \frac{1}{K} \sum_{\boldsymbol{x}_i \in \mathcal{N}_K} I(y_i = c)$$

for a training sample $\mathcal{D} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^N$.

K-nearest neighbours classifier

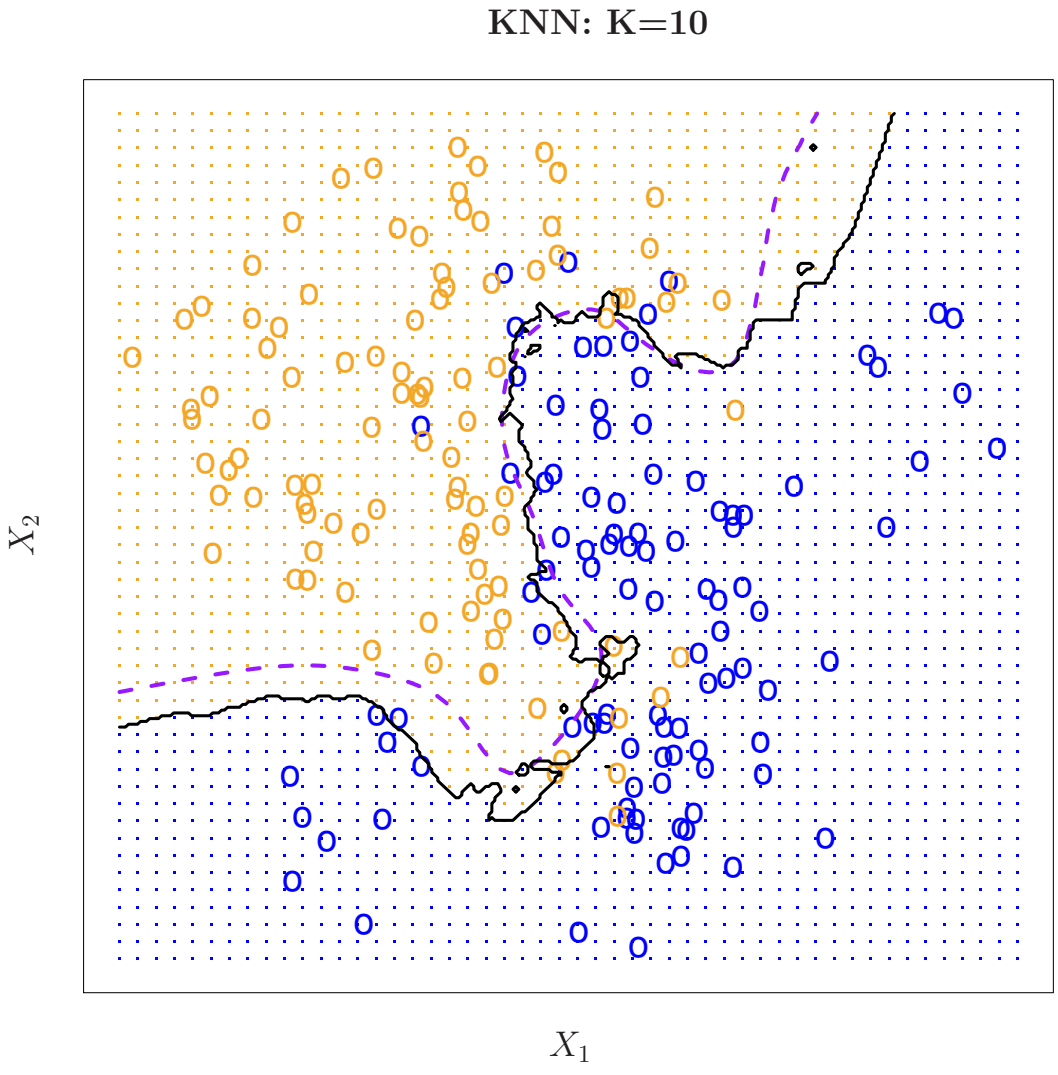




K-nearest neighbours classifier

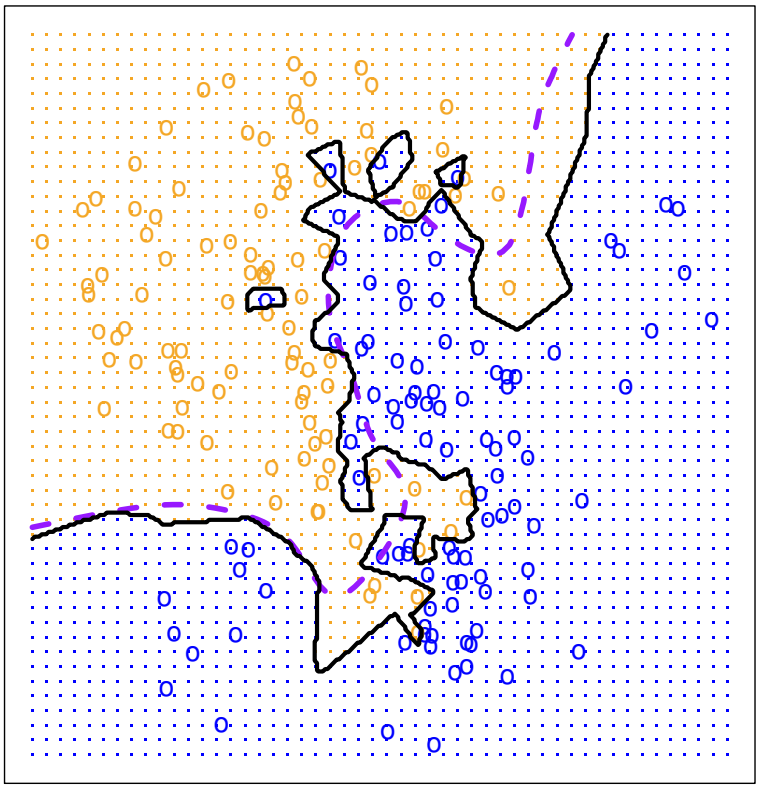
- In words, the KNN method finds the K training input points which are closest to $\boldsymbol{x}$, and computes the conditional probability as the fraction of those points that belongs to class c.
- Similarly to the KNN regression method, the KNN classifier is a direct nonparametric approximation to 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.

K-nearest neighbours classifier

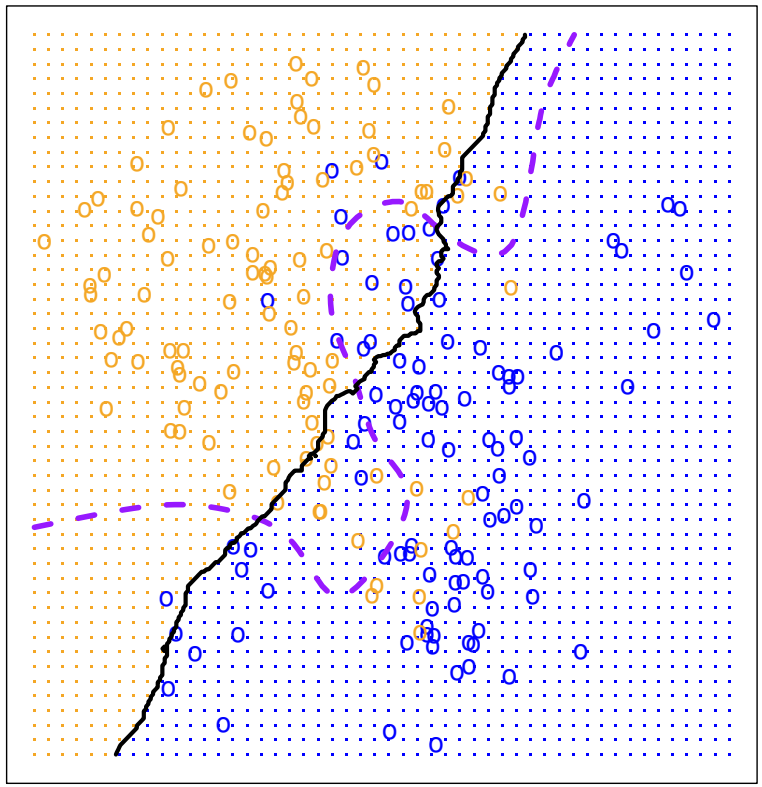


K-nearest neighbours classifier

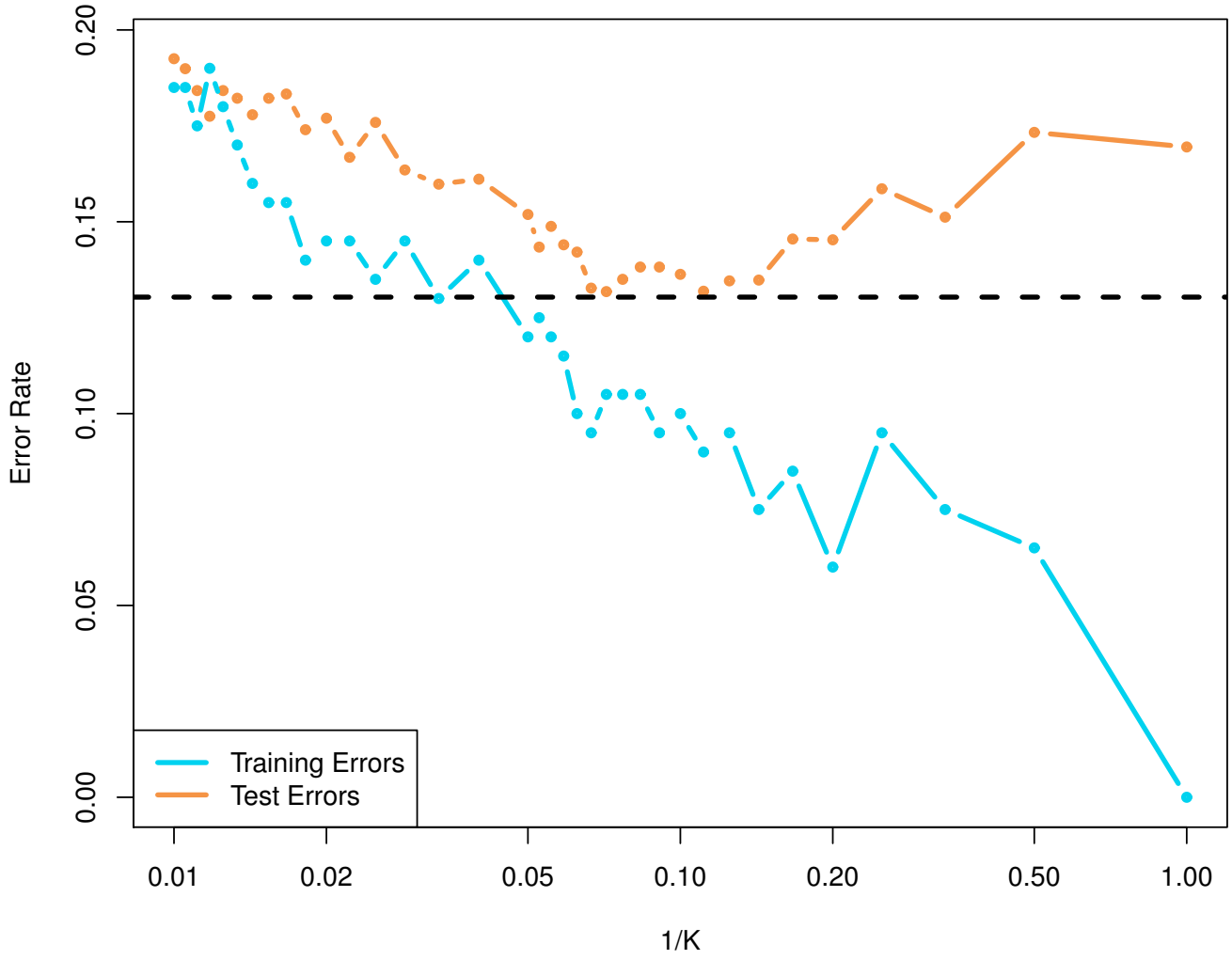
KNN: K=1



KNN: K=100



K-nearest neighbours classifier



Naïve Bayes classifier

Generative classifiers (key concept)

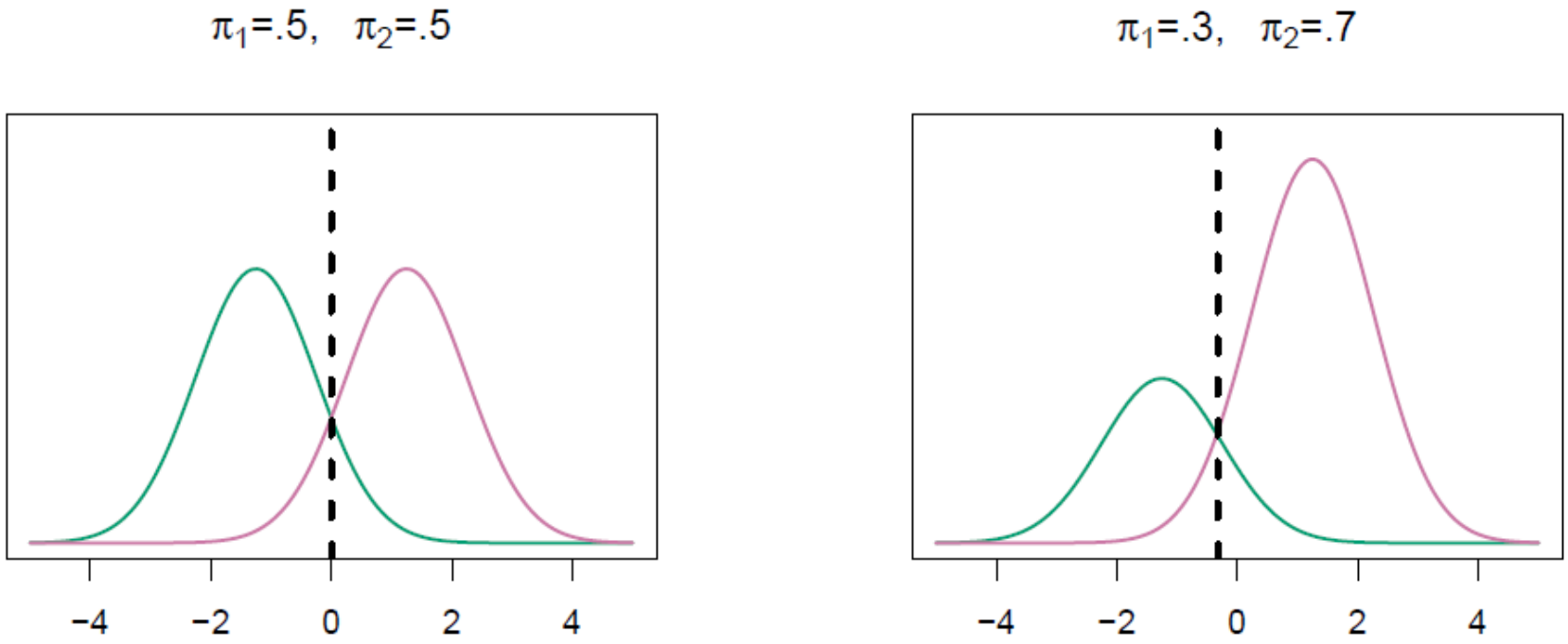
A **generative classifier** is a model that specifies how to generate the data given the **class conditional densities** $p(\boldsymbol{x}|y=c)$ and the (prior) class probabilities $p(y=c)$. This is a model for the joint distribution $p(y, \boldsymbol{x})$.

We compute the conditional probabilities for classification using Bayes' theorem,

$$p(y=c|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=c)p(y=c)}{\sum_{c' \in \mathcal{Y}} p(\boldsymbol{x}|y=c')p(y=c')}.$$

Generative classifiers

Consider the case with one normally distributed predictor and two classes. We classify to the highest input density, taking the prior into account.



Generative vs discriminative classifiers

In contrast, **discriminative classifiers** directly estimate the class conditional probabilities $p(y = c|\boldsymbol{x})$. Some examples are the KNN classifier, logistic regression, classification trees, support vector machines, and neural networks.

In a discriminative model, the distribution of the inputs is arbitrary.

We will return to the discussion of generative vs discriminative classifiers in the next module.

Naïve Bayes classifier (key concept)

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

The class conditional density the becomes

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

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.

Continuous predictors

For real valued predictors, we may assume that

$$X_j|Y = c \sim N(\mu_{jc}, \sigma^2_{jc}),$$

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

In the Gaussian case,

$$p(x_j|y = c) = \frac{1}{\sqrt{2\pi}\sigma_{jc}} \exp\left(-\frac{(x_j - \mu_{jc})^2}{2\sigma^2_{jc}}\right).$$

Continuous predictors

- Often, we first transform the predictors (for example by fitting Box-Cox transformations) in order to make the variables approximately normal or symmetric.

- We can also use other distributional assumptions or follow a nonparametric approach to estimate the class conditional densities by using kernel density estimates.

- We can also easily combine parametric and nonparametric univariate models for predictors.

Categorical predictors

When the predictors are binary, $x_j \in \{0, 1\}$, we use the Bernoulli distribution

$$X_j | Y = c \sim \text{Bernoulli}(\theta_{jc}),$$

where θ_{jc} is the probability that feature j occurs in class c.

In the case of categorical features, $x_j \in \{1, \ldots, K\}$ we use the multinoulli distribution

$$X_j | Y = c \sim \text{Cat}(\boldsymbol{\theta}_{jc}),$$

where $\boldsymbol{\theta}_{jc}$ is a K-vector of probabilities conditional on class c.

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 element records whether a word is present in the document or not. Hence, $x_{ij} = 1$ if word j appears in document i, and $x_{ij} = 0$ otherwise.

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

Application: document classification

The class conditional probability mass function is

$$p(\boldsymbol{x}_i|y_i = c, \boldsymbol{\theta}) = \prod_{j=1}^{p} \theta_{jc}^{x_{ij}} (1 - \theta_{jc})^{1-x_{ij}},$$

where p is the number of words in the **dictionary** and θ_{jc} is the probability that word j appears in a document of type c.

Maximum likelihood estimation

We usually estimate the Naïve Bayes model by maximum likelihood as follows:

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

2. Fit univariate models separately for each predictor within each class.

Maximum likelihood estimation

Let $P(Y = c) = \pi_c$. The joint probability for each observation is

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

The log-likelihood therefore is

$$L(\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{c=1}^{C} N_c \log(\pi_c) + \sum_{c=1}^{C} \sum_{j=1}^{p} I(y_i = c) \log\, p(x_{ij}; \boldsymbol{\theta}_{jc}),$$

where $N_c = \sum_{i=1}^{N} I(y_i = c)$ is the number of cases in category c.

Maximum likelihood estimation

The log-likelihood

$$L(\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{c=1}^{C} N_c \log(\pi_c) + \sum_{c=1}^{C} \sum_{j=1}^{p} I(y_i = c) \log\, p(x_{ij}; \boldsymbol{\theta}_{jc})$$

decomposes into several terms that concern different parameters. We can therefore optimise all these parameters separately.

For example, by maximising $\sum_{c=1}^{C} N_c \log(\pi_c)$, we obtain the estimates for the class priors

$$\widehat{\pi}_c = \frac{N_c}{N}.$$

Maximum likelihood estimation

Suppose that the predictors are binary, such that

$$X_j|Y = c \sim \text{Bernoulli}(\theta_{jc}).$$

We obtain the probability parameters as

$$\widehat{\theta}_{jc} = \underset{\theta_{jc}}{\text{argmax}} \sum_{i:\, y_i = c} \sum_{j=1}^{p} \log\, p(x_{ij}; \boldsymbol{\theta}_{jc}).$$

The MLE is

$$\widehat{\theta}_{jc} = \widehat{P}(X_j = 1|Y = c) = \frac{N_{jc}}{N_c},$$

where $N_{jc} = \sum_{i=1}^{N} I(x_{ij} = 1) I(y_i = c)$ is the number of cases in category c where predictor j is one.

Naïve Bayes classifier

- Despite being based on an assumption that is not 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 highly biased probabilities, the simplifications brought by it may lead to substantial savings in variance.

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 much lower than the loss in case of fraud.

Classification outcomes

We 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 (key concept)

The context will 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 for this business scenario. Our decision making should therefore take this into account.

Decision rule (key concept)

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

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

where τ is a decision threshold parameter.

Optimal decision (key concept)

Our decision problem is therefore to select an optimal threshold τ.

$$\tau^* = \underset{0<\tau<1}{\text{argmin}}\ E\left[L(Y, \delta_\tau(\boldsymbol{x}))|X = \boldsymbol{x}\right].$$

Let $\pi = P(Y = 1|X = \boldsymbol{x})$ to simplify the notation. We compare the expected loss from each decision,

$$E\left[L(Y, \delta_\tau(\boldsymbol{x}))|X = \boldsymbol{x}\right] = \begin{cases} \pi L_{\mathsf{TP}} + (1-\pi)L_{\mathsf{FP}} \text{ if } \delta_\tau(\boldsymbol{x}) = 1, \\ \pi L_{\mathsf{FN}} + (1-\pi)L_{\mathsf{TN}} \text{ if } \delta_\tau(\boldsymbol{x}) = 0. \end{cases}$$

Optimal decision (key concept)

The optimal decision threshold corresponds to the probability value π such that the loss from a positive or negative classification is equal.

$$\tau^* L_{\mathsf{TP}} + (1 - \tau^*) L_{\mathsf{FP}} = \tau^* L_{\mathsf{FN}} + (1 - \tau^*) L_{\mathsf{TN}}$$

Therefore, the optimal threshold 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 $L_{\mathsf{TP}} = L_{\mathsf{TN}} = 0$. Therefore,

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

Example: credit scoring

In the credit scoring example, we have that $L_{\mathsf{TP}} = -L_{\mathsf{FN}}$ (profit equals missed profit) and $L_{\mathsf{TN}} = -L_{\mathsf{FP}}$ (avoided default loss equals default loss).

Therefore,

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

Example: credit scoring

Optimal threshold for loan decision:

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

We expect that the loss from default to be much higher than the profit from a loan to a creditworthy customer ($L_{\mathsf{FP}} >> L_{\mathsf{FN}}$), leading to a high threshold τ^*.

It is only worth it to lend to customers that have a high probability of repayment.

Model evaluation for binary classification

Confusion matrix (key concept)

A **confusion matrix** counts the number of true negatives, false positives, false negatives, and true positives for the test data.

		Classification		
		$\widehat{Y} = 0$	$\widehat{Y} = 1$	Total
Actual	$Y = 0$	True negatives (TN)	False positives (FP)	N
	$Y = 1$	False negatives (FN)	True positives (TP)	P
	Total	Negative predictions	Positive predictions	

Estimating the generalisation error

- Estimating the generalisation error is straightforward using the loss and confusion matrices.
- As always, it is important to quantify the uncertainty in the estimate by reporting the standard error or doing interval estimation.
- For the rest of this section, we discuss important concepts for assessing binary classification models that do not rely on the loss matrix.

Sensitivity and specificity (key concepts)

The **sensitivity**, recall, or true positive rate is

$$P(\widehat{Y} = 1|Y = 1) = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{True positives}}{\text{Actual positives}}.$$

The **specificity** is

$$P(\widehat{Y} = 0|Y = 0) = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{True negatives}}{\text{Actual negatives}}.$$

False positive and false negative rates

The **false positive rate** (FPR) is

$$P(\widehat{Y} = 1|Y = 0) = \frac{\text{FP}}{\text{TN} + \text{FP}} = \frac{\text{False positives}}{\text{Actual negatives}} = 1 - \text{Specificity}.$$

The **false negative rate** (FNR) is

$$P(\widehat{Y} = 0|Y = 1) = \frac{\text{FN}}{\text{TP} + \text{FN}} = \frac{\text{False negatives}}{\text{Actual positives}} = 1 - \text{Sensitivity}.$$

Trade-off between sensitivity and specificity (key concept)

- There is a trade-off between sensitivity and specificity, since a classifier can obtain maximum sensitivity (specificity) by setting $\tau = 1$ ($\tau = 0$) and always returning positive (negative).

- Equivalently, this is a trade-off between sensitivity and achieving a lower false positive rate.

Example: credit scoring

- Decreasing the threshold makes the loan decisions more lenient, leading to loans to customers with lower probability of full repayment.

- Issuing additional loans will increase both the number of true positives (higher sensitivity) and false positives (lower specificity).

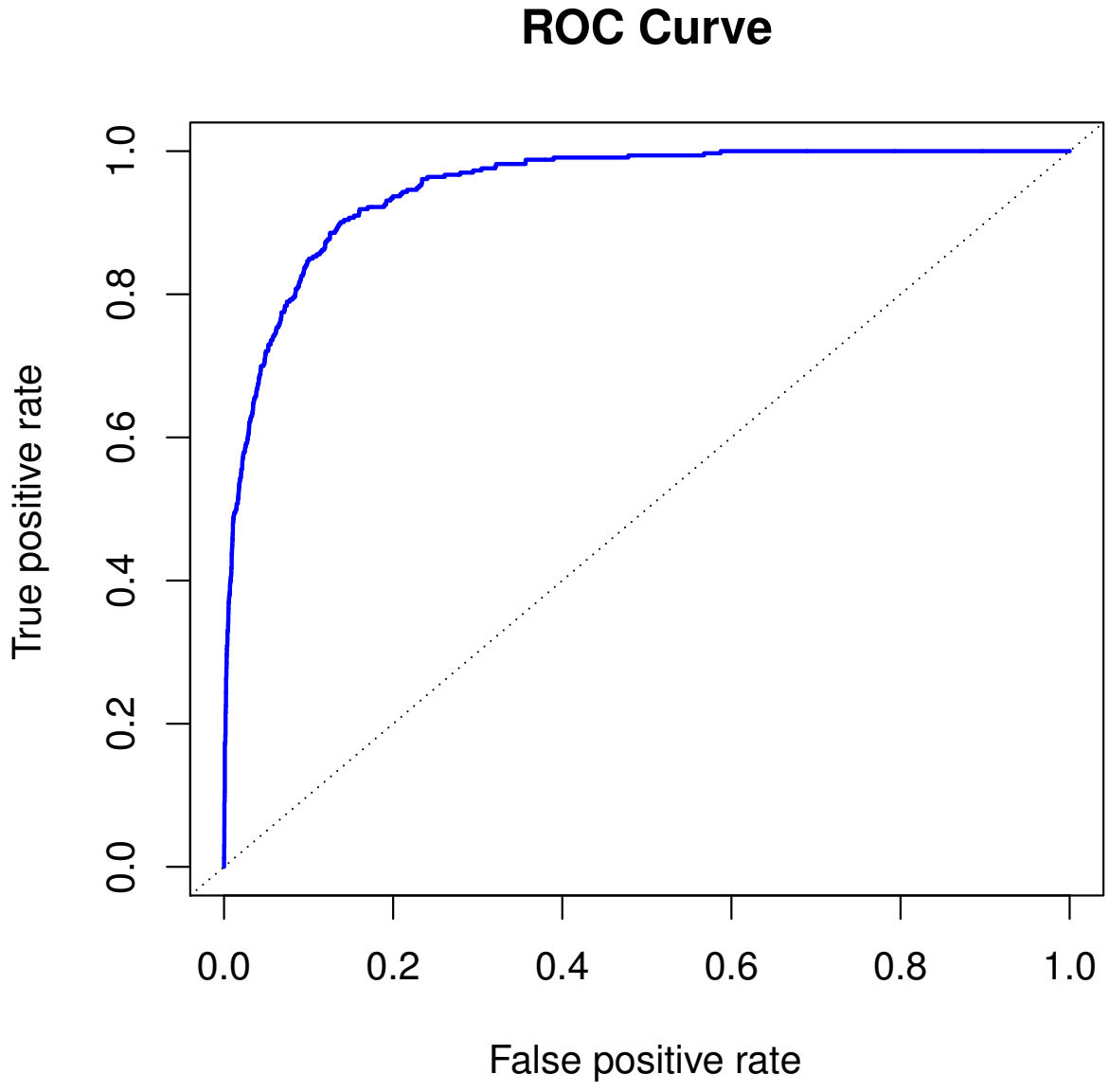
ROC curve (key concept)

A **receiver operating characteristic** or **ROC** curve plots the sensitivity against specificity or the false positive rate for a range of threshold values τ.

We can read the the ROC plot as telling us the false positive rate that we need to accept to obtain a given level of sensitivity.

We often summarise the quality of ROC curve as a single number using the **area under the curve** or **AUC**. Higher AUC scores are better, with a maximum of one.

ROC curve


ROC Curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Imbalanced classes

Many classification scenarios (such as fraud detection) concern rare events, leading to a very large proportion of negatives in the data.

In this situation we say that the classes are highly **imbalanced**.

The specificity is not very informative for these problems, as it will tend to be high regardless of the quality of the classifier (nearly all transactions are legitimate and classified as such).

Precision (key concept)

In the imbalanced scenario, we are usually more interested in the proportion of detections that are actually positive. We define the **precision** as

$$P(Y = 1|\widehat{Y} = 1) = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{True positives}}{\text{Positive classifications}}$$

The **false discovery rate** (FDR) is one minus the precision.

$$P(Y = 0|\widehat{Y} = 1) = \frac{\text{FP}}{\text{TP} + \text{FP}} = \frac{\text{False positives}}{\text{Positive classifications}}$$

Precision recall curve

A **precision recall curve** plots the precision against the recall (sensitivity) as we vary the threshold τ. The mean precision (averaging over recall values) approximates the area under the precision recall curve.

Example: transaction fraud detection

- In fraud detection, the bank is concerned that too many false alarms (low precision) would lead to high costs of investigating flagged transactions.
- The bank will weigh this cost against the savings from catching fraudulent transactions.
- Therefore, the financial institution is primarily interested in the precision recall curve.
- Increasing τ reduces the number of false alarms.

Review questions

- What is classification?
- What is a zero-one loss?
- What is the Bayes classifiers?
- What is the misclassification rate?
- Explain the KNN classifier.
- What is a loss matrix?

Review questions

- How do we formulate a decision rule for binary classification?
- What is a confusion matrix? Write down what the matrix looks like.
- What are sensitivity, specificity, and precision?
- Why is there a trade-off between sensitivity and specificity?
- What is the difference between discriminative and generative classifiers?
- What is the key assumption of the Naive Bayes classifier?