Lecture 7

Probabilistic Models

M. Poel
G. Englebienne

University of Twente

- Introduction
- Parametric Models
  - Categorical variables
  - Continuous Variables
- Expectation and Expected loss
  - Decision threshold
- Simplifying assumptions
- Incorporating prior information
  - Maximum a Posteriori
  - Bayesian learning
- Wrap-up

Introduction

Parametric Models
- Categorical variables
- Continuous Variables

Expectation and Expected loss
- Decision threshold

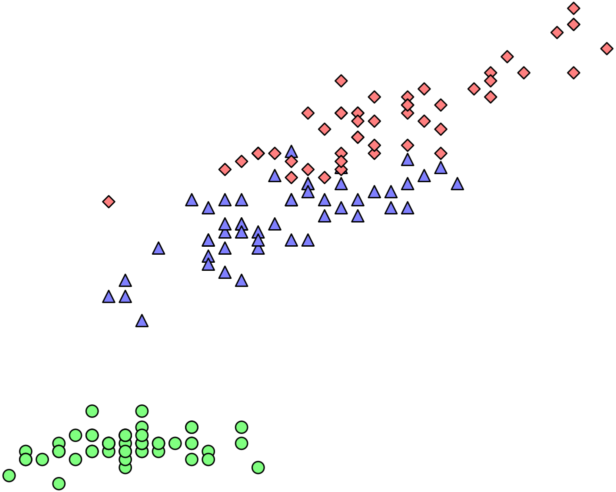
Simplifying assumptions

Incorporating prior information
- Maximum a Posteriori
- Bayesian learning

Wrap-up

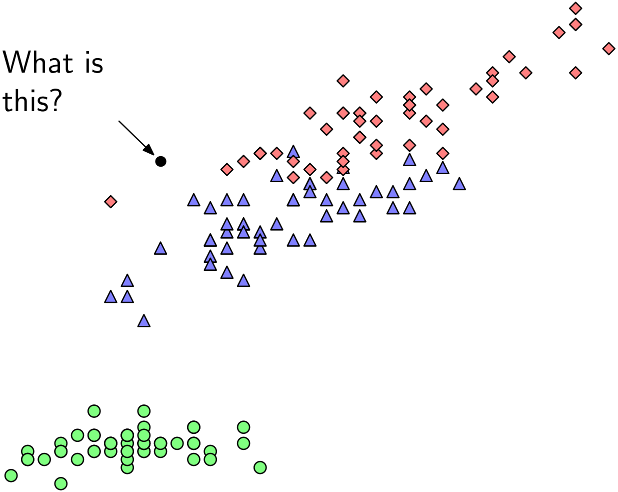
Probabilities and Distributions

Example (Iris Data)



Probabilities and Distributions

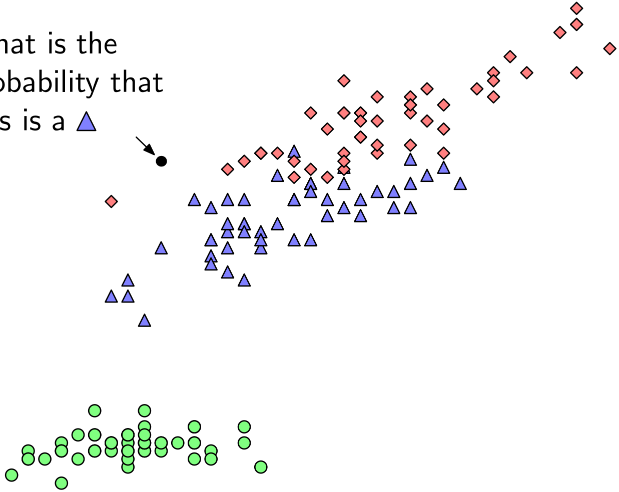
Example (Iris Data)



What is this?

Probabilities and Distributions

Example (Iris Data)



What is the probability that this is a △

Probabilistic Modelling

We want to know

$$p(\mathcal{C}|\mathbf{x})$$

Two possibilities:

1. Discrimintative Models: learn a $p(\mathcal{C}|\mathbf{x})$ directly

$$p(\mathcal{C}|\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta})$$

   For example, logistic regression: $p(\mathcal{C}|\mathbf{x}) = \sigma(\mathbf{x}, \mathbf{w})$

2. Generative model: learn the class-wise distribution of your data, and use Bayes' rule

$$p(\mathcal{C}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C})\, p(\mathcal{C})}{p(\mathbf{x})}$$

Bayes' Theorem

$$\overset{\text{Posterior}}{p(\mathcal{C}|\mathbf{x})} = \frac{\overset{\text{Likelihood}}{p(\mathbf{x}|\mathcal{C})}\;\overset{\text{Prior}}{p(\mathcal{C})}}{\underset{\text{Evidence}}{p(\mathbf{x})}}$$

Questions:

1. How do we compute $p(\mathbf{x}|\mathcal{C})$?
2. How do we use $p(\mathcal{C}|\mathbf{x})$ to make decisions?

Bayes' Theorem

$$\overset{\text{Posterior}}{p(\mathcal{C}|\mathbf{x})} = \frac{\overset{\text{Likelihood}}{p(\mathbf{x}|\mathcal{C})}\,\overset{\text{Prior}}{p(\mathcal{C})}}{\underset{\text{Evidence}}{p(\mathbf{x})}}$$

Questions:

1. How do we compute $p(\mathbf{x}|\mathcal{C})$?
2. How do we use $p(\mathcal{C}|\mathbf{x})$ to make decisions?

- Introduction
- Parametric Models
  - Categorical variables
  - Continuous Variables
- Expectation and Expected loss
  - Decision threshold
- Simplifying assumptions
- Incorporating prior information
  - Maximum a Posteriori
  - Bayesian learning
- Wrap-up

Parametric estimation for Categorical Variables

Example (Biased Coins)

Consider two otherwise undistinguishable coins; coin A is fair, coin B is biased and has a probability of 0.7 of landing on its head. You pick a coin at random and throw the following sequence:

Tail Head Tail Head Tail Head Head

What is the probability that this is coin A?

Parametric estimation for Categorical Variables

Example (Biased Coins)

First, we notice that the throws do not affect each other:

$$\begin{aligned} p(t,h,t,h,t,h,h) &= p(t)\,p(h)\,p(t)\,p(h)\,p(t)\,p(h)\,p(h) \\ &= p(t)^3 p(h)^4 \end{aligned}$$

The throws depend on which coin we used, and we picked one of the two coins with equal probability, so we can use Bayes' theorem to compute the posterior probability that coin B has been used:

$$\begin{aligned} p(C=a|X=\{t,h,t,h,t,h,h\}) &= \frac{p(t,h,t,h,t,h,h|a)\,p(a)}{p(t,h,t,h,t,h,h)} \\ &= \frac{0.5^3\,0.5^4\,0.5}{0.5^3\,0.5^4\,0.5 + 0.3^3\,0.7^4\,0.5} \\ &\approx 0.55 > 0.5 \end{aligned}$$

The Bernoulli Distribution

The Bernoulli distribution is the probability mass function of a binary event $x \in \{0, 1\}$:

$$p(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

This leads to the following log-likelihood of a set of i.i.d. measurements:

$$\begin{aligned}\mathcal{L}(x_{1\ldots N}; \mu) &= \log \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n} \\ &= \sum_{n=1}^{N} x_n \log \mu + (1 - x_n) \log(1 - \mu)\end{aligned}$$

The Maximum-Likelihood estimator for the parameter μ is found by maximising the log-likelihood

MLE for Bernoulli

We find the maximum of the function by setting its first derivative to zero:

$$
\begin{aligned}
\frac{\partial}{\partial\mu}\mathcal{L}(x_{1\ldots N};\mu) &= \frac{\partial}{\partial\mu}\sum_{n=1}^{N} x_n \log\mu + (1-x_n)\log(1-\mu) \\
0 &= \sum_{n=1}^{N} \frac{x_n}{\mu} - \frac{1-x_n}{1-\mu} = \sum_{n=1}^{N} \frac{x_n\,(1-\mu) - (1-x_n)\,\mu}{\mu\,(1-\mu)} \\
0 &= \sum_{n=1}^{N} x_n - \mu\,\mathbf{x}_n - \mu + \mu\,x_n = \sum_{n=1}^{N} x_n - \mu \\
\mu &= \frac{\sum_{n=1}^{N} x_n}{N}
\end{aligned}
$$

The Categorical Distribution

The Bernouli distribution can be extended to multiple classes

- Using one-of-K encoding ($x = k \in \{1, \ldots, K\} \Rightarrow \mathbf{x} : x_{j \neq k} = 0, x_k = 1$)

$$p(\mathbf{x}) = \prod_{i=1}^{K} p_i^{x_i} \qquad \text{where} \sum_{i=1}^{K} p_i = 1$$

- The MLE estimator for p_i is given by:

$$p_i = \frac{\sum_{n=1}^{N} x_i^{(n)}}{N}$$

Dealing with Continuous Variables

Example (Waiting for the bus)

What is the probability that you will have to wait *exactly* 5 minutes for your bus?

Dealing with Continuous Variables

Example (Waiting for the bus)

What is the probability that you will have to wait *exactly* 5 minutes for your bus?

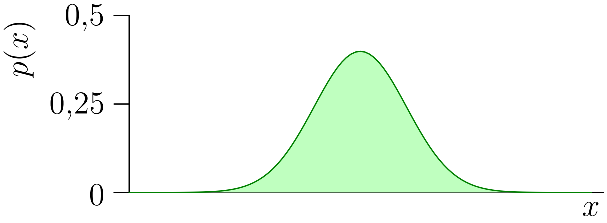
zero.

We cannot assign probabilities to values. Instead, we need to assign probabilities to value *ranges*

Example (A better question would be:)

What is the probability that you will have to wait between 5 and 6 minutes for your bus?

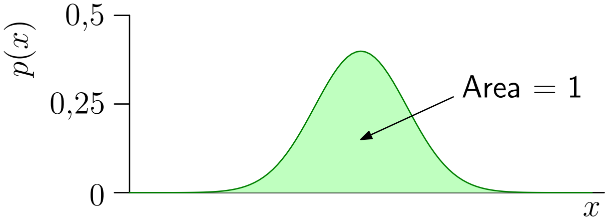
Probability Density Functions



Probability Density Functions (PDFs):

- Are non-negative everywhere, integrate to 1
- Not probabilities, can be interpreted as *relative* probabilities
  - Can be larger than 1

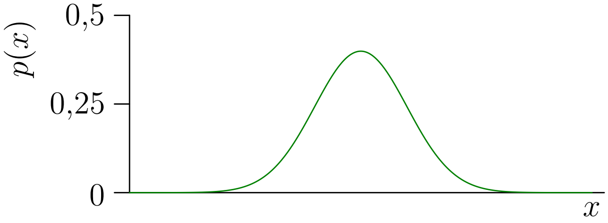
Probability Density Functions



Probability Density Functions (PDFs):

- Are non-negative everywhere, integrate to 1
- Not probabilities, can be interpreted as *relative* probabilities
  - Can be larger than 1

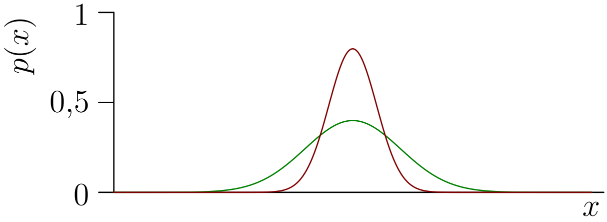
Probability Density Functions



Probability Density Functions (PDFs):

- Are non-negative everywhere, integrate to 1
- Not probabilities, can be interpreted as *relative* probabilities
  - Can be larger than 1

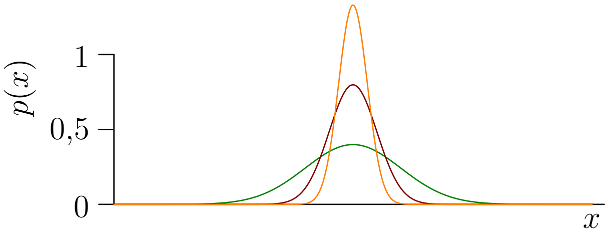
Probability Density Functions



Probability Density Functions (PDFs):

- Are non-negative everywhere, integrate to 1
- Not probabilities, can be interpreted as *relative* probabilities
    - Can be larger than 1

Probability Density Functions

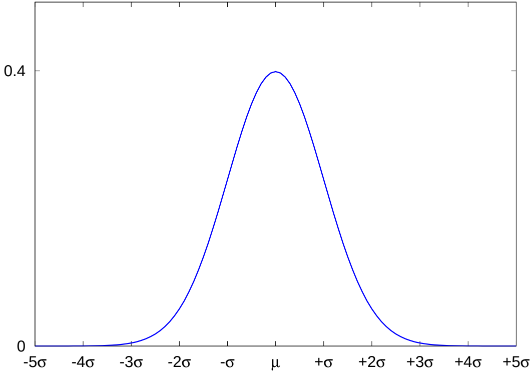


Probability Density Functions (PDFs):

- Are non-negative everywhere, integrate to 1
- Not probabilities, can be interpreted as *relative* probabilities
  - Can be larger than 1

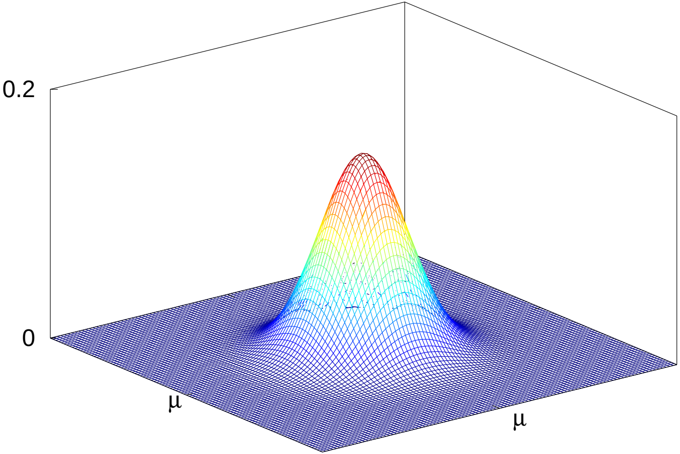
The Gaussian or Normal distribution

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp-\frac{(x-\mu)^2}{2\sigma^2} \tag{1}$$



The Gaussian or Normal distribution

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/1}|\boldsymbol{\Sigma}|^{1/2}} \exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{2}$$



Central Limit Theorem

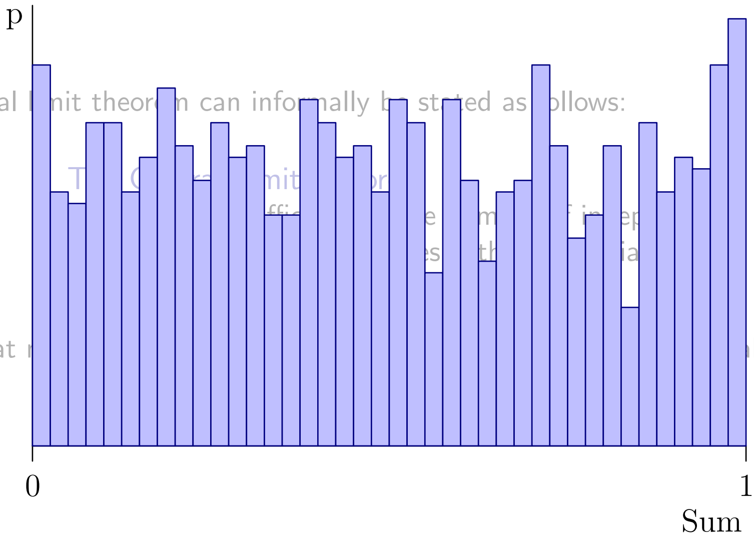
The central limit theorem can informally be stated as follows:

The Central Limit Theorem

The sum of a sufficiently large number of independent, identically distributed variables with finite variance will have an approximately Gaussian distribution.

Notice that no assumption is made about the distribution of these variables

Central Limit Theorem



The central limit theorem can informally be stated as follows:

The Central Limit Theor

ffic e m f indep

es th ia

Notice that r ables

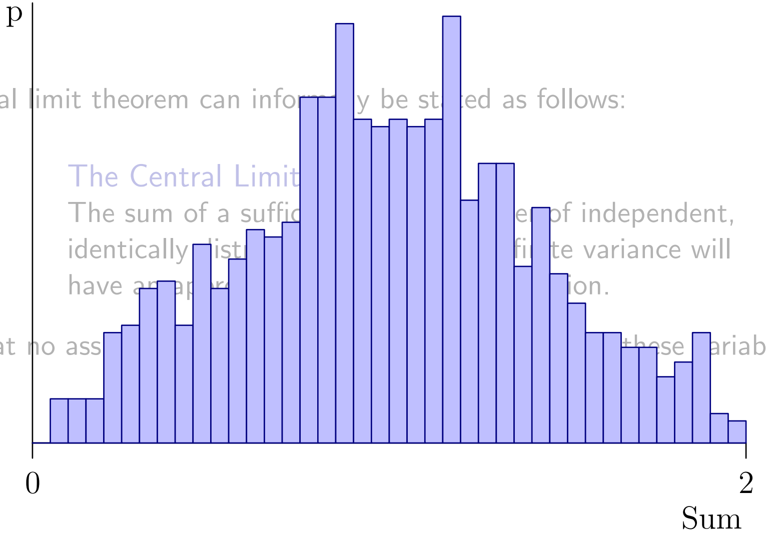
Central Limit Theorem

The central limit theorem can infor y be st ted as follows:

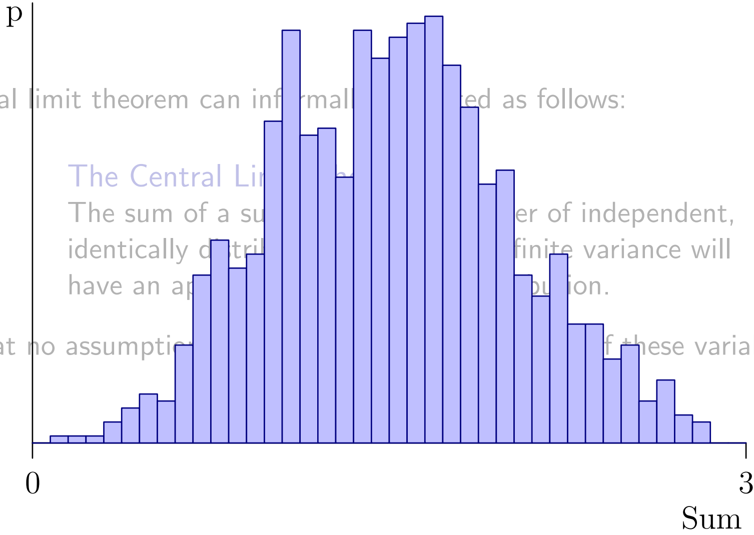
The Central Limit

The sum of a suffic er of independent, identically istr fi te variance will have ar ap r ion.

Notice that no ass these ariables



Central Limit Theorem



The central limit theorem can inf rmall ted as follows:

The Central Li

The sum of a su er of independent, identically d stri finite variance will have an ap ion.

Notice that no assumptio f these variables

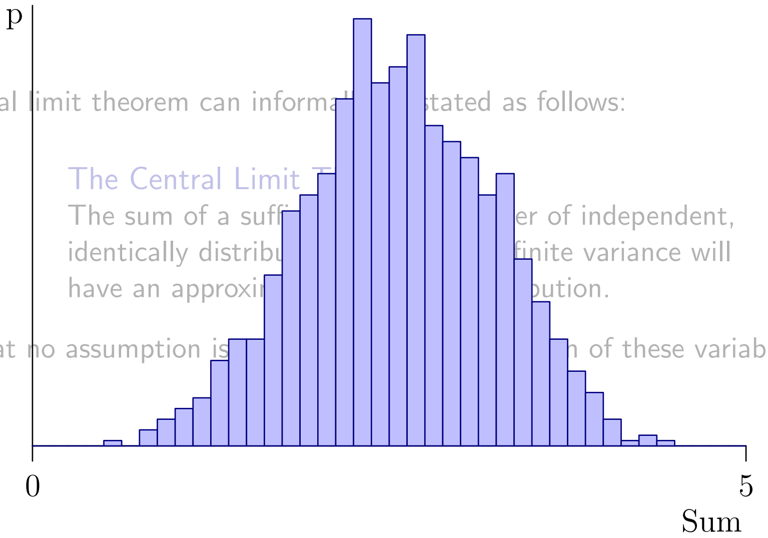
Central Limit Theorem

The central limit theorem can informa stated as follows:

The Central Limit T

The sum of a suffi er of independent,
identically distribu finite variance will
have an approxi bution.

Notice that no assumption is n of these variables



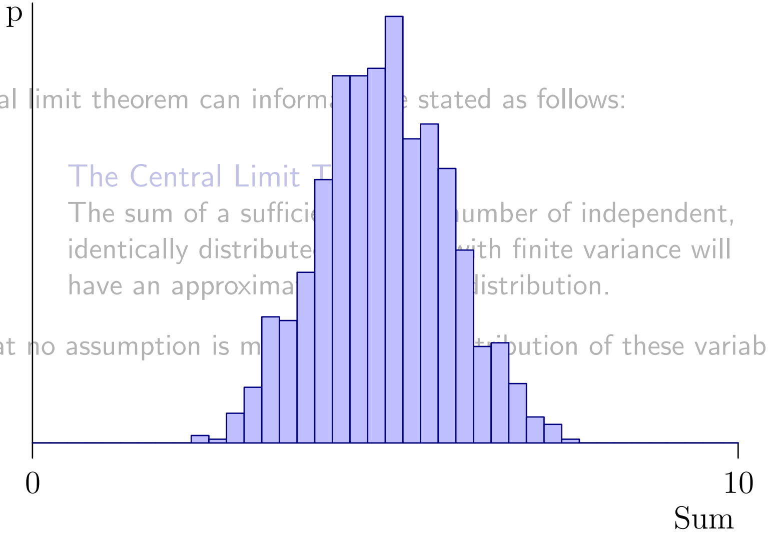
Central Limit Theorem

The central limit theorem can informa e stated as follows:

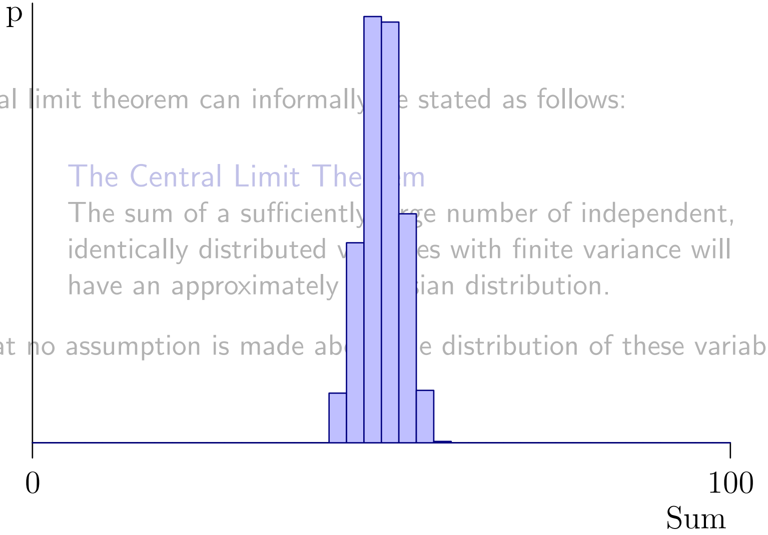
The Central Limit T

The sum of a suffici umber of independent, identically distribute ith finite variance will have an approxima distribution.

Notice that no assumption is m ibution of these variables



Central Limit Theorem



The central limit theorem can informally be stated as follows:

The Central Limit Theorem

The sum of a sufficiently large number of independent, identically distributed variables with finite variance will have an approximately Gaussian distribution.

Notice that no assumption is made about the distribution of these variables

Central Limit Theorem

The central limit theorem can informally be stated as fol

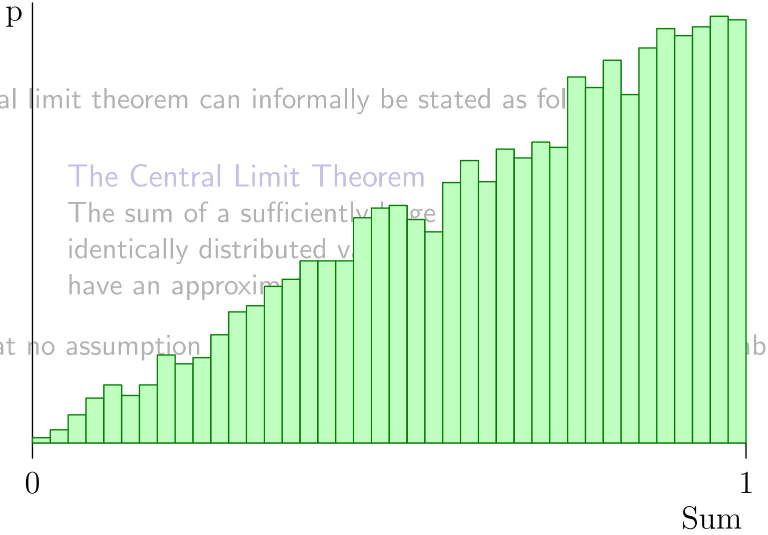
The Central Limit Theorem

The sum of a sufficiently l ge

identically distributed v

have an approxim

Notice that no assumption ables



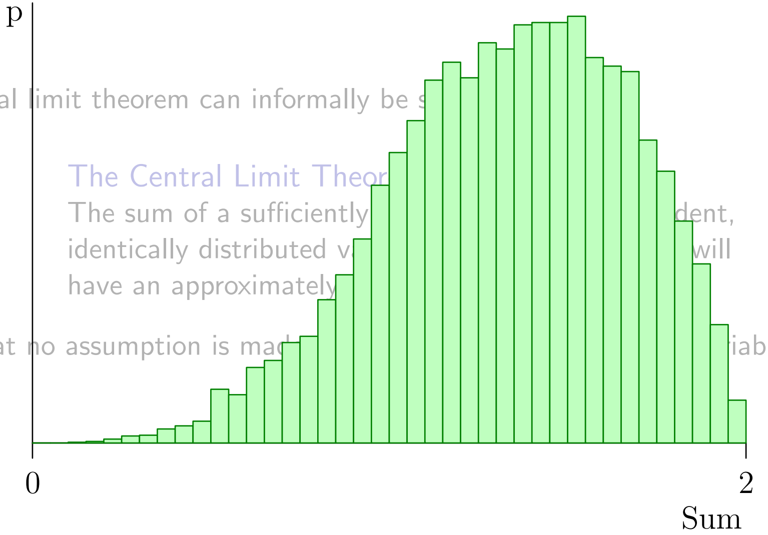
Central Limit Theorem

The central limit theorem can informally be s

The Central Limit Theor

The sum of a sufficiently dent, identically distributed v will have an approximately

Notice that no assumption is mad iables



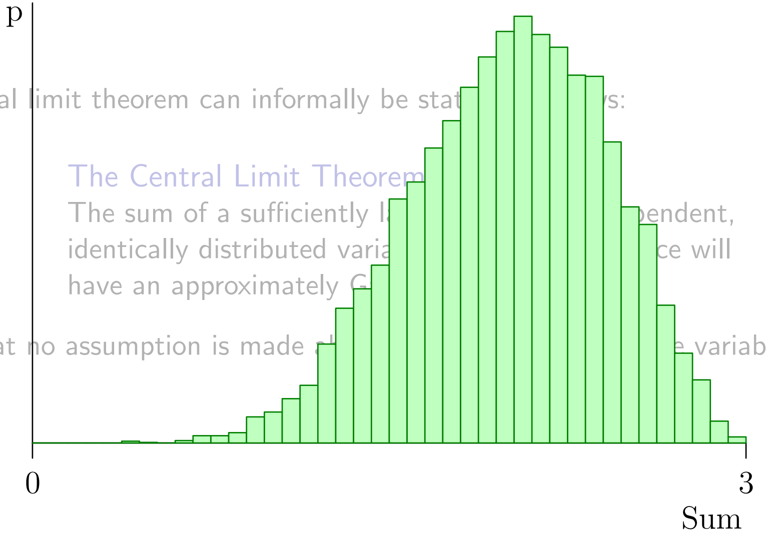
Central Limit Theorem

The central limit theorem can informally be stat s:

The Central Limit Theorem

The sum of a sufficiently l pendent, identically distributed varia ce will have an approximately G

Notice that no assumption is made a e variables



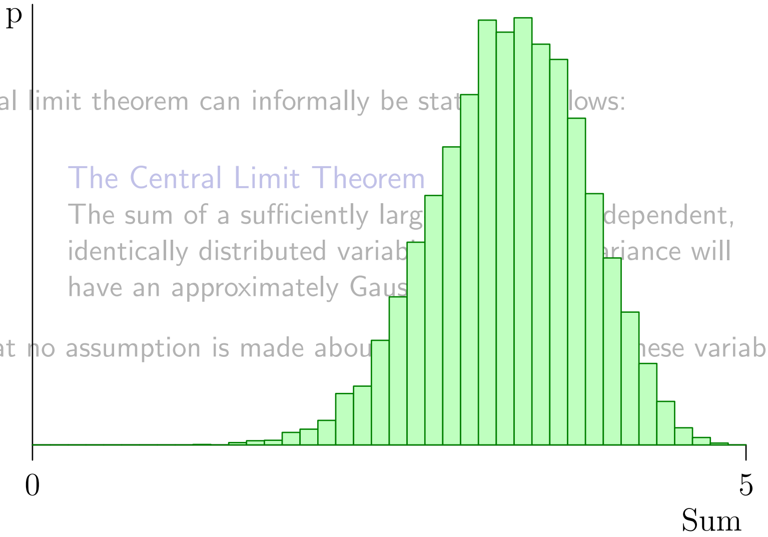
Central Limit Theorem

The central limit theorem can informally be stat lows:

The Central Limit Theorem

The sum of a sufficiently larg dependent, identically distributed variab ariance will have an approximately Gaus

Notice that no assumption is made abou hese variables



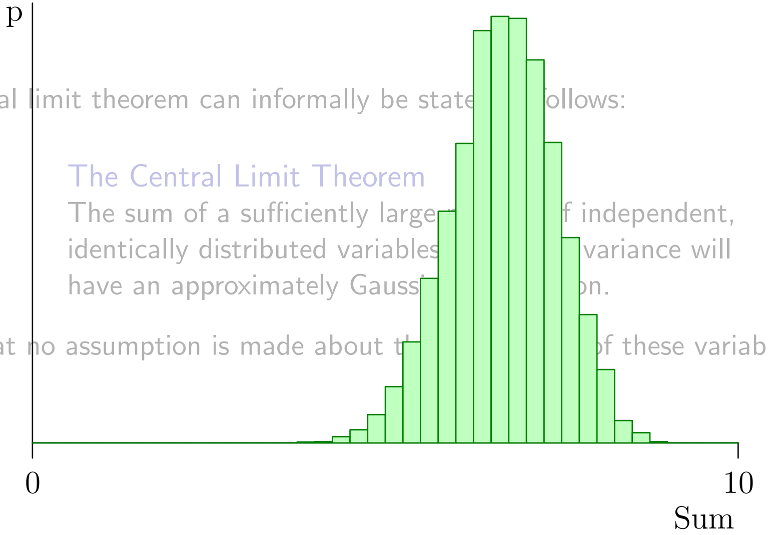
Central Limit Theorem

The central limit theorem can informally be state ollows:

The Central Limit Theorem

The sum of a sufficiently large f independent, identically distributed variables variance will have an approximately Gauss on.

Notice that no assumption is made about t f these variables



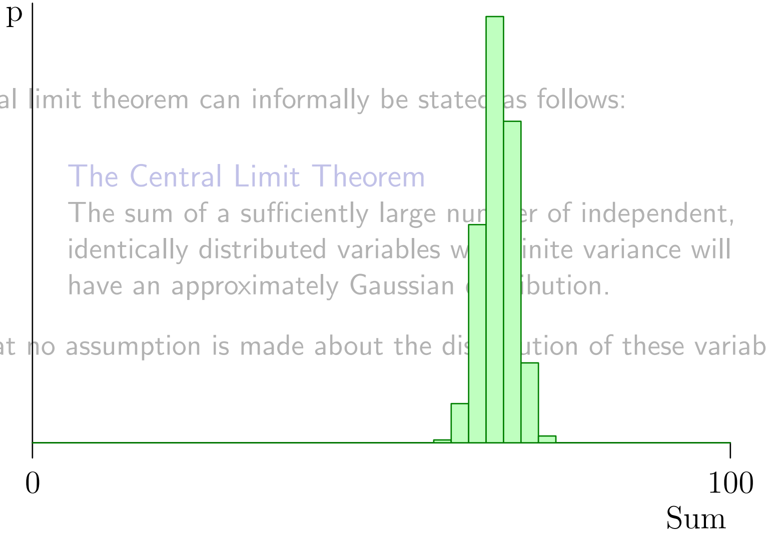
Central Limit Theorem

The central limit theorem can informally be stated as follows:

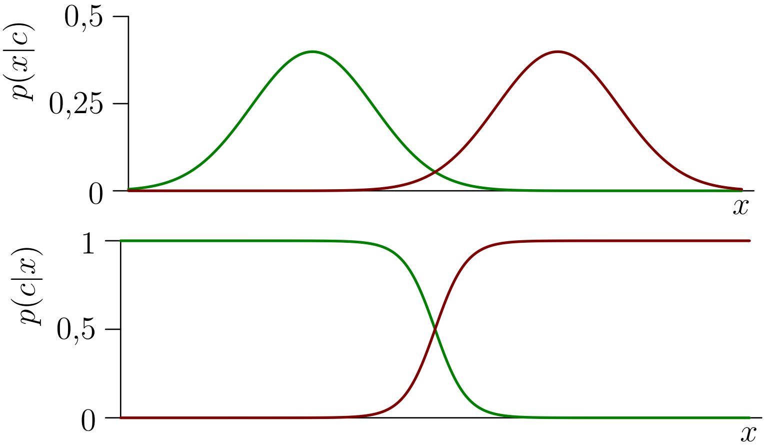
The Central Limit Theorem

The sum of a sufficiently large number of independent, identically distributed variables with finite variance will have an approximately Gaussian distribution.

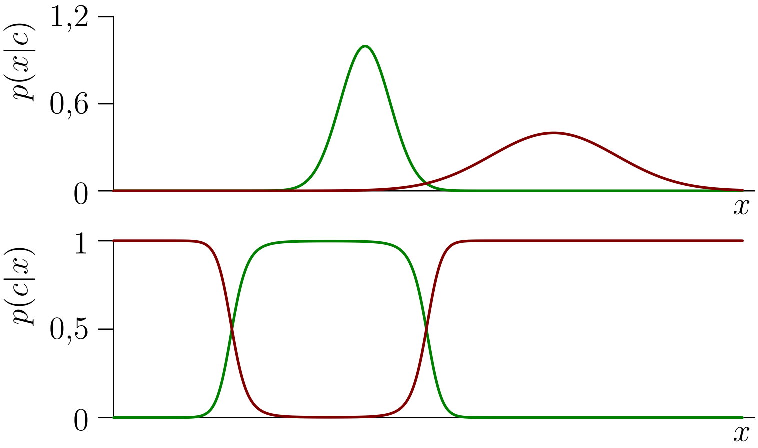
Notice that no assumption is made about the distribution of these variables



Bayes' Theorem



Bayes' Theorem



Introduction

Parametric Models

- Categorical variables
- Continuous Variables

Expectation and Expected loss

- Decision threshold

Simplifying assumptions

Incorporating prior information

- Maximum a Posteriori
- Bayesian learning

Wrap-up

Expectation

The expectation of a variable (or function of that variable)

$$\mathbb{E}_{p(x)}[f(\mathbf{x})] = \int f(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

is the value that the long-term average of the function converges to.

- From our perspective, it's our "best bet" value
- When making decision, we want to minimise our expected error

Decision threshold

Classification: obtain a feature vector **x** and predict the corresponding class $\mathcal{C}$

Example

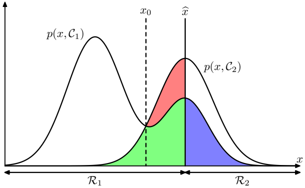
Given an X-ray image, predict the health state of the person

Bayesian decision rule

- Assign class labels so as to minimise the expected error
- Assign an observation **x** to class $\mathcal{C}_i$ if

$$p(\mathcal{C}_i|\mathbf{x}) > p(\mathcal{C}_j|\mathbf{x}) \qquad \forall j \neq i \tag{3}$$

Minimising the misclassification rate



From Bayes' rule, we have that

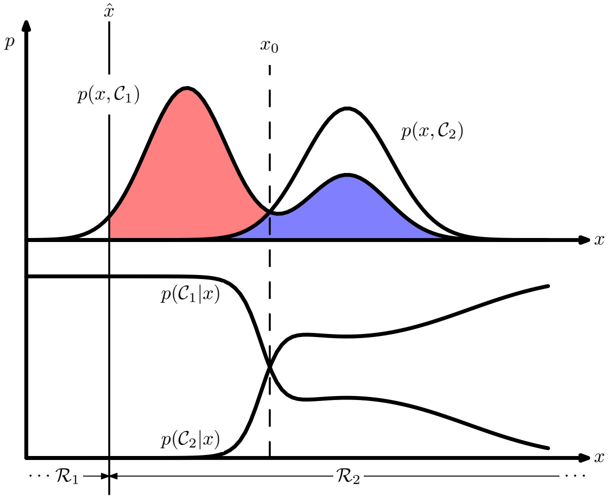
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} \tag{4}$$

We want to minimise the probability of a mistake, that is:

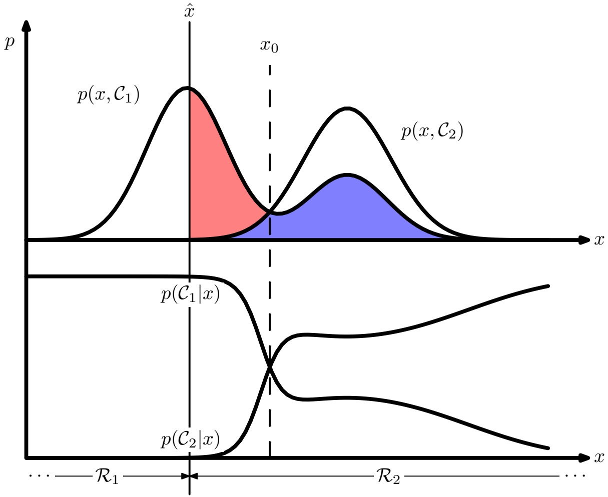
$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) && (5) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) && (6) \end{aligned}$$

Since $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$ and $p(\mathbf{x})$ is the same in both terms, $p(\text{mistake})$ is minimal if each point $\mathbf{x}$ is assigned to the class for which $p(\mathcal{C}_k|\mathbf{x})$ is largest.

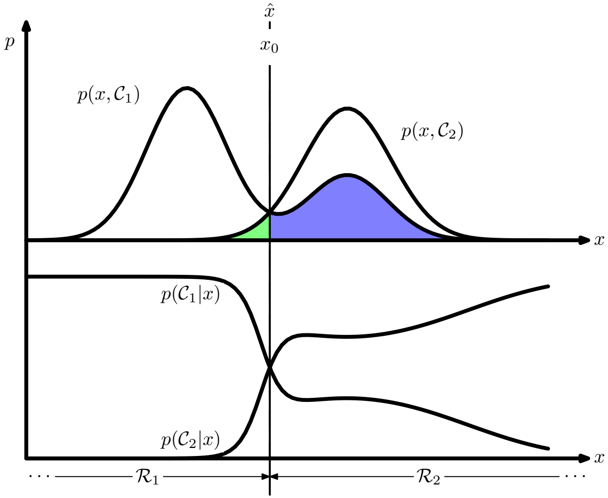
Minimising the misclassification rate



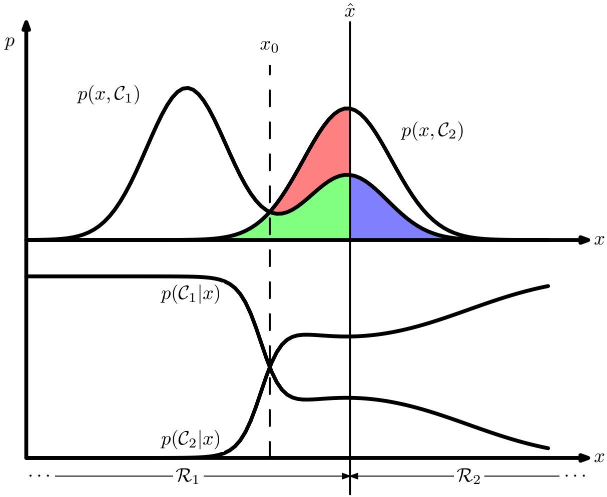
Minimising the misclassification rate



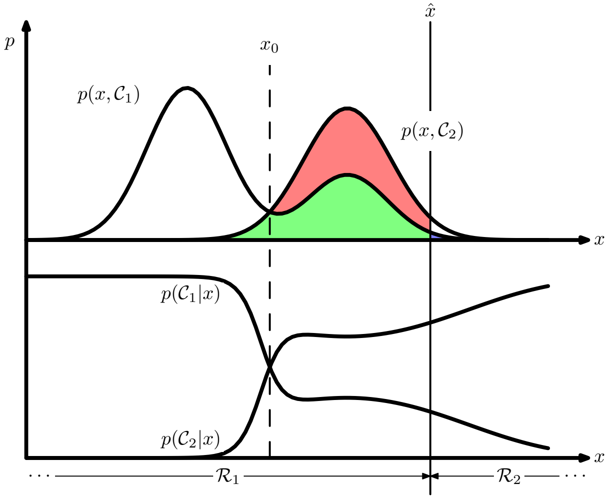
Minimising the misclassification rate



Minimising the misclassification rate



Minimising the misclassification rate



Reject option

In some cases, the posterior probability $p(\mathcal{C}_k|\mathbf{x})$ of the most likely class may be far less than one.

- The regions where this is the case lead to most misclassifications

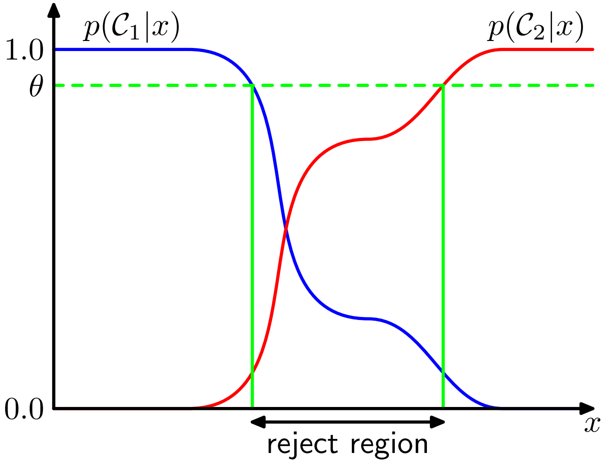
In some cases it is better to avoid making a decision when that is the case, in order to improve the performance on the examples for which a decision is made.

Example

In medical image classification, it may be suitable to automatically classify images for which we are very confident and leave the difficult cases for a human to evaluate.

The reject option

Achieved by choosing a threshold, $\boldsymbol{\theta}$, and rejecting datapoints for which the largest $p(\mathcal{C}_k|\mathbf{x}) \leqslant \boldsymbol{\theta}$.



Minimising the expected loss

In the case of unbalanced misclassification costs: loss matrix

Cancer classification example

$$L = \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \tag{7}$$

The expected loss is then given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \tag{8}$$

which is minimised by assigning each datapoint $\mathbf{x}$ to the class j for which

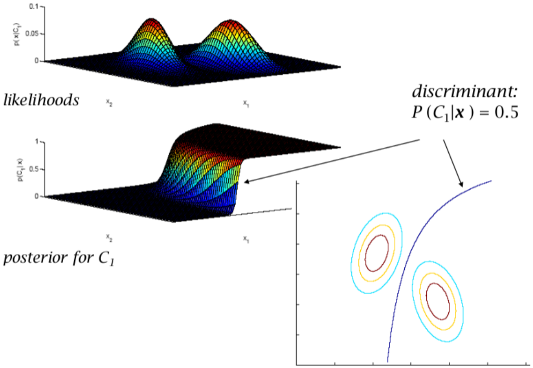
$$\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x}) \tag{9}$$

- Introduction
- Parametric Models
  - Categorical variables
  - Continuous Variables
- Expectation and Expected loss
  - Decision threshold
- Simplifying assumptions
- Incorporating prior information
  - Maximum a Posteriori
  - Bayesian learning
- Wrap-up

Simplifying Assumptions

We often lack data to train a model well by MLE

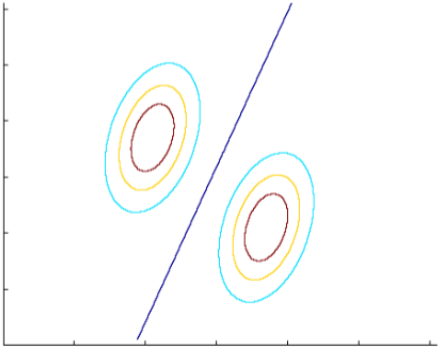
- To address this, we can simplify the model
- Reduce the number of parameters without losing too much discriminating power
- Examples (Gaussian PDF):
    - Assume classes share the same covariance
    - Assume the covariances are diagonal
    - Assume that covariance matrices are spherical



Simplifying Assumptions

We often lack data to train a model well by MLE

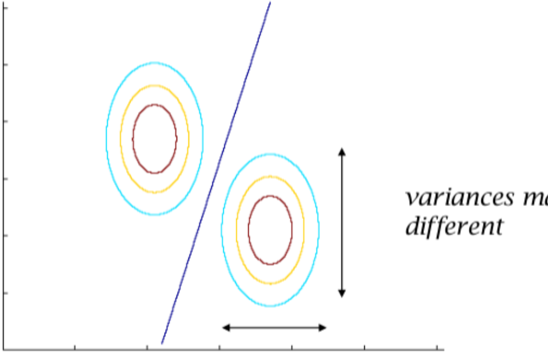
- To address this, we can simplify the model
- Reduce the number of parameters without losing too much discriminating power
- Examples (Gaussian PDF):
  - Assume classes share the same covariance
  - Assume the covariances are diagonal
  - Assume that covariance matrices are spherical



Simplifying Assumptions

We often lack data to train a model well by MLE

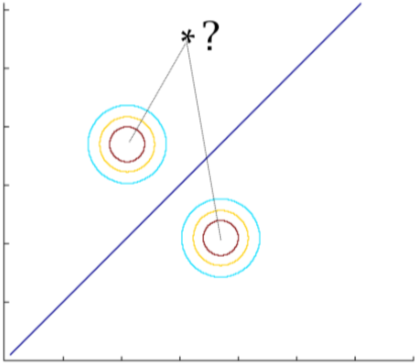
- To address this, we can simplify the model
- Reduce the number of parameters without losing too much discriminating power
- Examples (Gaussian PDF):
  - Assume classes share the same covariance
  - Assume the covariances are diagonal
  - Assume that covariance matrices are spherical



Simplifying Assumptions

We often lack data to train a model well by MLE

- To address this, we can simplify the model
- Reduce the number of parameters without losing too much discriminating power
- Examples (Gaussian PDF):
  - Assume classes share the same covariance
  - Assume the covariances are diagonal
  - Assume that covariance matrices are spherical



Naive Bayes

Naive Bayes: Assume all data dimensions are independent given the class

$$\begin{aligned} p(\mathbf{x}|\mathcal{C}) &= p(x_1|\mathcal{C})\cdots p(x_N|\mathcal{C}) \\ &= \prod_i p(x_i|\mathcal{C}) \end{aligned}$$

Features:

- Scales linearly in the number of features
- Overly confident if features are not independent
- Performs surprisingly well in practice
- Beware: Naive Bayes is not “Bayesian Learning”
- Notice: conditional independence $\neq$ marginal independence

$$p(x_1,\ldots,x_n) = \sum_{\mathcal{C}} p(x_1|\mathcal{C})\cdots p(x_N|\mathcal{C})p(\mathcal{C}) \neq p(x_1)\cdots p(x_N)$$

Naive Bayes

Naive Bayes: Assume all data dimensions are independent given the class

$$\begin{aligned} p(\mathbf{x}|\mathcal{C}) &= p(x_1|\mathcal{C}) \cdots p(x_N|\mathcal{C}) \\ &= \prod_i p(x_i|\mathcal{C}) \end{aligned}$$

Features:

- Scales linearly in the number of features
- Overly confident if features are not independent
- Performs surprisingly well in practice
- Beware: Naive Bayes is not “Bayesian Learning”
- Notice: conditional independence $\neq$ marginal independence

$$p(x_1, \ldots, x_n) = \sum_{\mathcal{C}} p(x_1|\mathcal{C}) \cdots p(x_N|\mathcal{C}) p(\mathcal{C}) \neq p(x_1) \cdots p(x_N)$$

Introduction

Parametric Models

- Categorical variables
- Continuous Variables

Expectation and Expected loss

- Decision threshold

Simplifying assumptions

Incorporating prior information

- Maximum a Posteriori
- Bayesian learning

Wrap-up

Maximum A Posteriori (MAP) learning

Instead of learning the parameters that maximise the likelihood, why not learn the most likely parameters? Using Bayes' rule, we have:

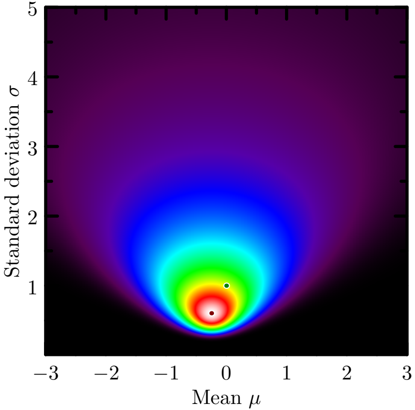
$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x})d\boldsymbol{\theta}} \tag{10}$$

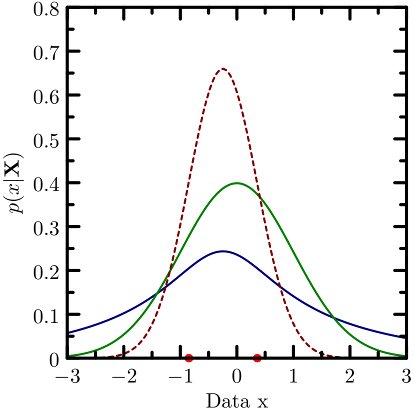
This requires us to place a prior over the parameter values

- Any prior is possible, choose prior to reflect prior knowledge
- If we use a Gaussian distribution with zero mean, this is equivalent to ML learning with parameter shrinkage
- The denominator is often intractable to compute but is constant, so that

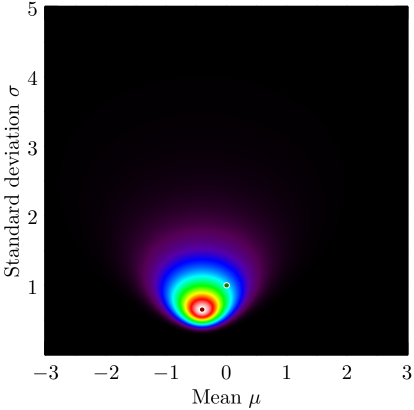
$$\arg\max_{\boldsymbol{\theta}} \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x})d\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{11}$$

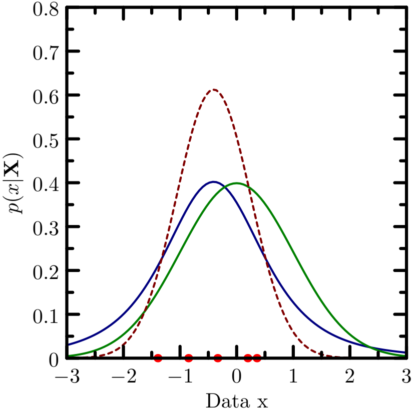
The Bayesian approach



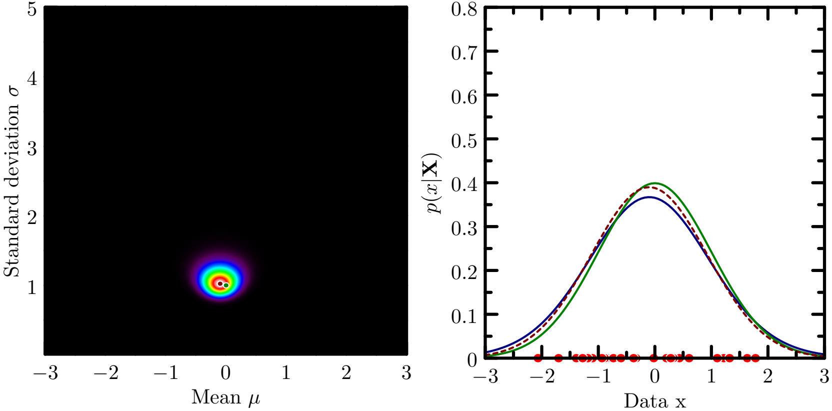


The Bayesian approach

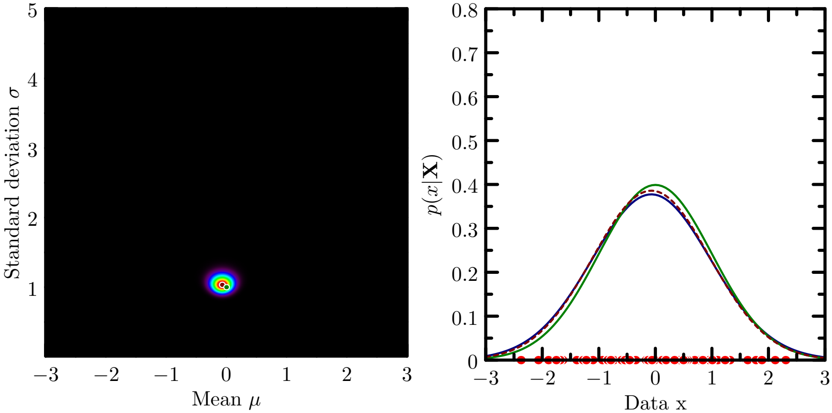




The Bayesian approach



The Bayesian approach



The Bayesian approach

In fact, we're not really interested in knowing the original distribution that "generated" the data

- We'll never know that anyway

What we really want to do, is to use the knowledge that we have in an optimal way. That is, we want

$$p(t|\mathbf{x},\mathbf{X},\mathbf{t}) = \int p(t|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{X},\mathbf{t})\mathrm{d}\boldsymbol{\theta} \tag{12}$$

In effect, we consider all the models (of the form that we have chosen beforehand) that could have generated the data, and weigh their prediction according to how probable they are.

Introduction

Parametric Models

- Categorical variables
- Continuous Variables

Expectation and Expected loss

- Decision threshold

Simplifying assumptions

Incorporating prior information

- Maximum a Posteriori
- Bayesian learning

Wrap-up

Summary

We've talked about Parametric Modelling

- Generative vs. Discriminative Models
- Probability Mass and Density Functions
- Learning the parameters of said functions
- Effect of Independence assumptions