Image Searching and Modelling using Machine Learning Methods

Part IB Paper 8
Information Engineering Elective

Lecture 1: Feature vectors and models

Zoubin Ghahramani

zoubin@eng.cam.ac.uk

Department of Engineering
University of Cambridge

Easter Term

What will we cover in part C ?

Image searching and modelling using machine learning methods

We will focus on the application of pattern recognition and statistical machine learning methods to ***image retrieval*** *and related problems. Although all examples will use images, the ideas are generally applicable to other domains, for example, web document retrieval, music, and financial data.*

Topics:

- Representing images as feature vectors
- Probabilistic models, use of Bayes rule, Bernoulli distributions and multivariate Gaussians
- Image retrieval
- Outlier removal and novelty detection
- A case study of an image retrieval method

Images

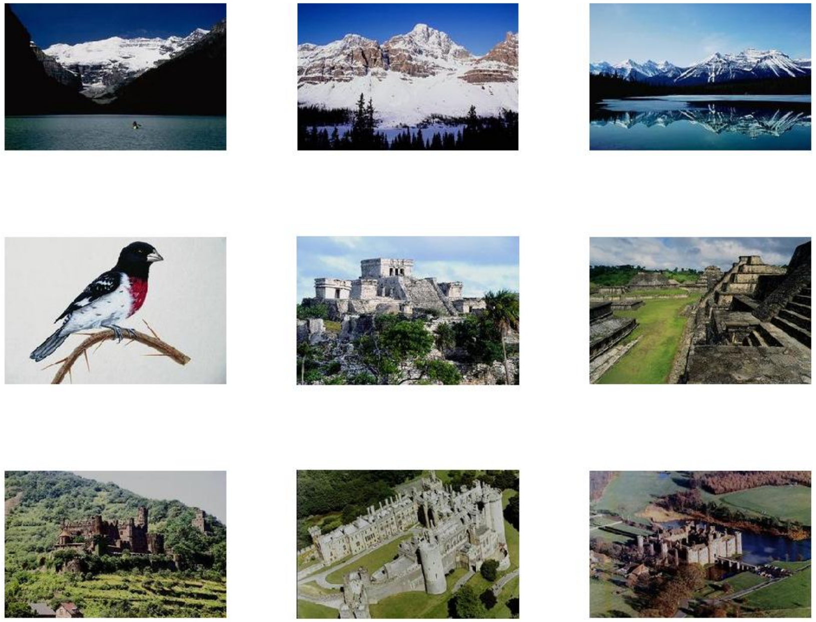


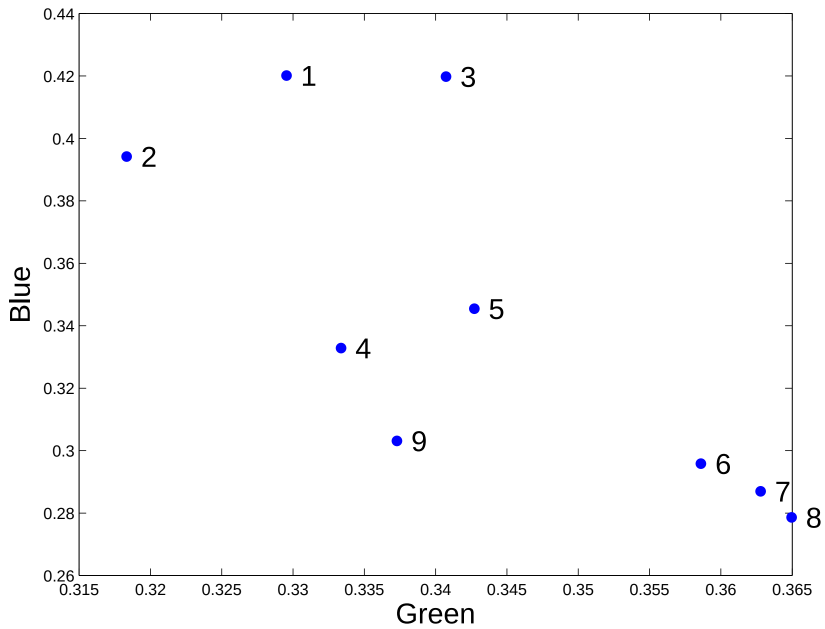


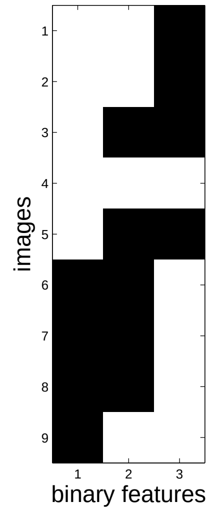




Representing Images as Feature Vectors







There are many possible feature vector represenions, e.g.:

- $\mathbf{x} = [r \;\; g \;\; b]$ overall red/green/blue values
- $\mathbf{x} = [p_1, \ldots p_N]$ vector of greyscale pixel values
- $\mathbf{x} = [w_1, \ldots, w_M]$ visual words

Different Types of Features

Let $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ denote D features of an image (or any other data object!).
Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_N\}$ be a data set of images

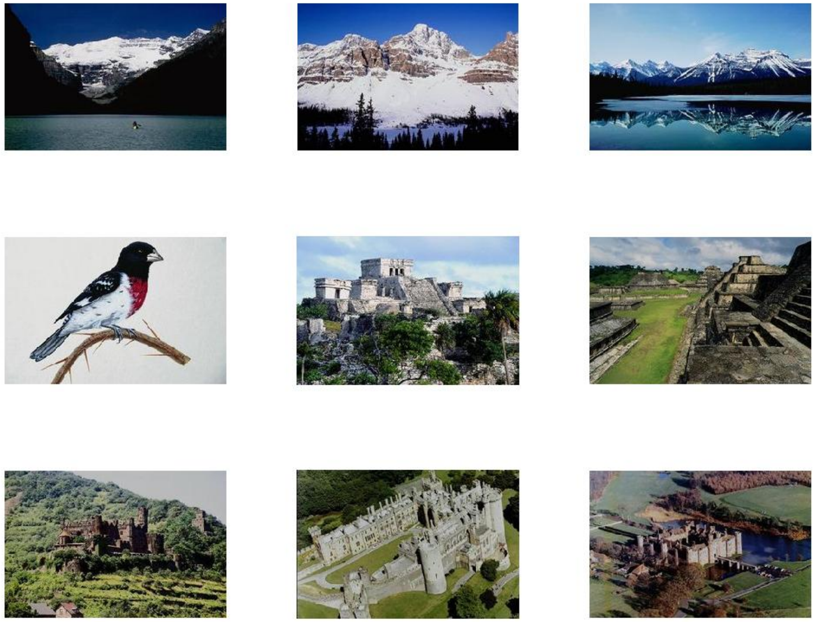
There are many possible types of features, e.g.:

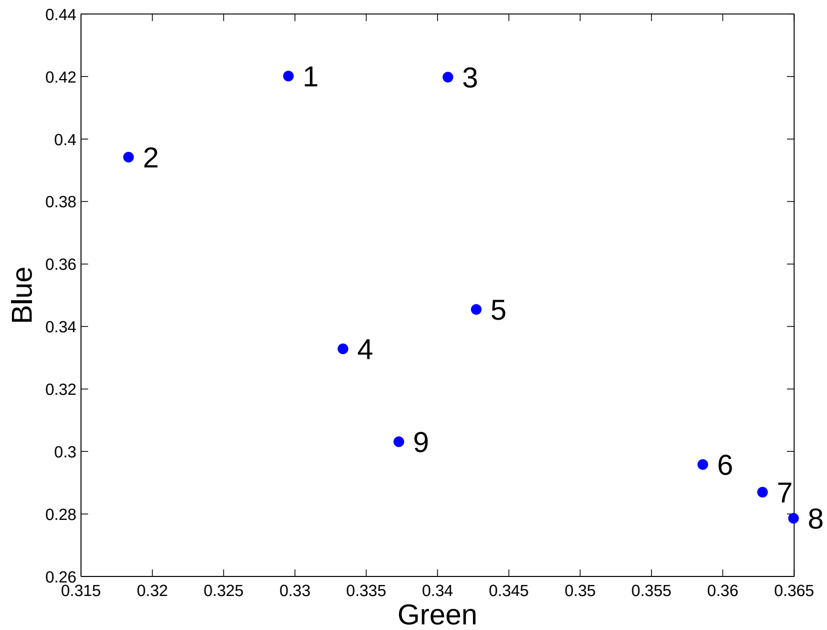
- $x_i \in \{0, 1\}$ - binary features
- $x_i \in \mathbb{R}$ - real-valued features
- $x_i \in \mathbb{R}_+$ - non-negative features
- $x_i \in \{0, 1, 2, \ldots\}$ - ordinal integer counts
- $x_i \in \{\text{cloud}, \text{sky}, \text{tree}, \ldots\}$ - nominal, categorical

Q: What can we do with feature vectors?
Q: How can we model them?

(We'll focus on binary and real-valued features)

What can we do with feature vectors?





- classification
- outlier removal
- modelling/prediction/completion
- retrieval

Binary Features

$$x_i \in \{0, 1\}$$

- A deterministic model does not represent uncertainty (e.g. leaves are green).
- A probabilistic model tries to capture the variability in the features (e.g. leaves are generally green)

For binary data:

$$P(x_i = 1) = \theta$$

where θ is the probability that feature i is 1, and

$$P(x_i = 0) = 1 - \theta$$

since x_i has to be either 0 or 1 for binary features.

The above two statements imply:

$$P(x_i) = \theta^{x_i}(1 - \theta)^{1 - x_i}$$

which is the Bernoulli distribution.

Multivariate Bernoulli

Univariate:

$$P(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Multivariate:

$$P(\mathbf{x}) = \prod_{i=1}^{D} \theta_i^{x_i}(1-\theta_i)^{1-x_i}$$

Q: What does θ_i represent?

Q: What is a limitation of this model?

To make the dependence of this model on its parameters explicit, we can write: $P(\mathbf{x}|\boldsymbol{\theta})$.

Comparing Data Points

Given a model parametrized by $\boldsymbol{\theta}$, and two data points $\mathbf{x}$ and $\mathbf{x}'$, we can find out which is more probable under the model.

$$r = \frac{P(\mathbf{x}|\boldsymbol{\theta})}{P(\mathbf{x}'|\boldsymbol{\theta})} > 1$$

means that $\mathbf{x}$ is more probable than $\mathbf{x}'$, given $\boldsymbol{\theta}$. Equivalently,

$$\log r = \log P(\mathbf{x}|\boldsymbol{\theta}) - \log P(\mathbf{x}'|\boldsymbol{\theta}) > 0$$

For example, for multivariate Bernoulli model:

$$\begin{aligned}
\log r &= \sum_{i=1}^{D} (x_i - x_i') \log \theta_i + (x_i' - x_i) \log(1 - \theta_i) \\
&= \sum_{i=1}^{D} (x_i - x_i') \log \frac{\theta_i}{1 - \theta_i}
\end{aligned}$$

Comparing Models

Given a data point or set of data points we can find out which of two parameters $\boldsymbol{\theta}$ or $\boldsymbol{\theta}'$ has higher likelihood

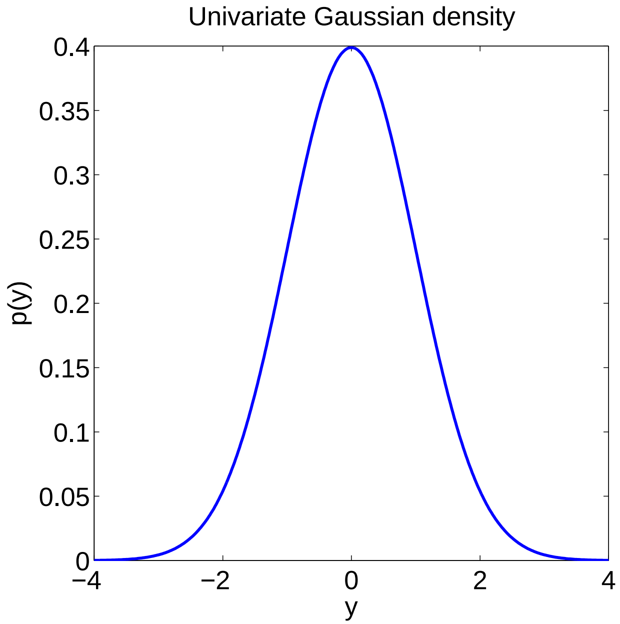
$$r = \frac{P(\mathbf{x}|\boldsymbol{\theta})}{P(\mathbf{x}|\boldsymbol{\theta}')}$$

Univariate Gaussians

$$x_i \in \mathbb{R}$$

Univariate Gaussian density $(x \in \mathbb{R})$:

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$



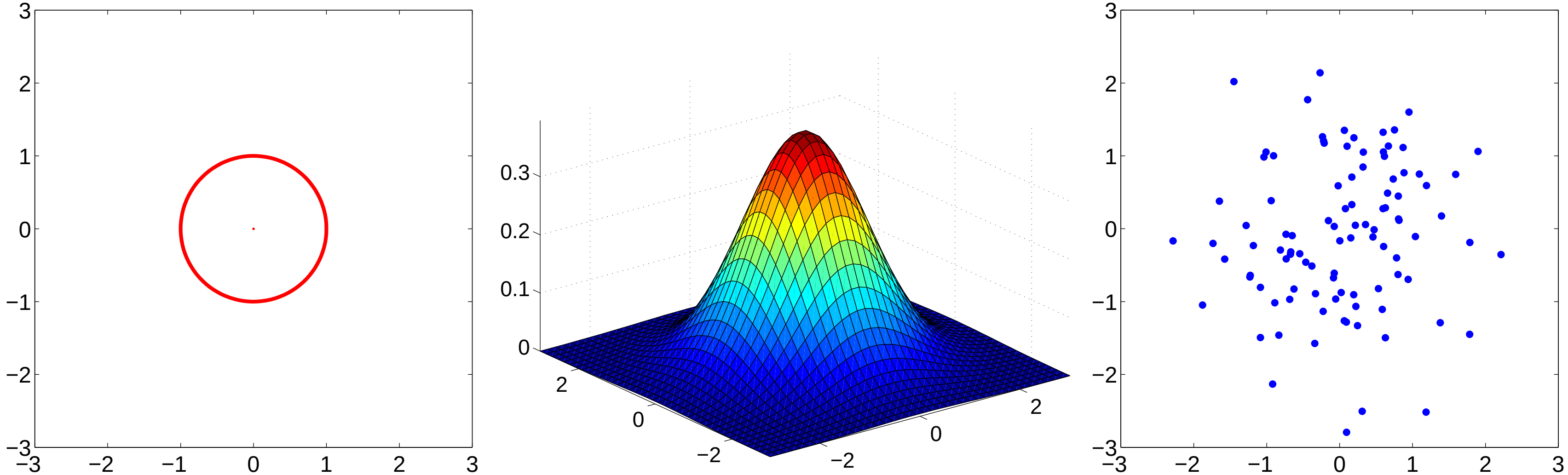
This model has parameters $\boldsymbol{\theta} = \{\mu, \sigma\}$ which model the mean and standard deviation of the data, respectively.

The multivariate Gaussian

Multivariate Gaussian density ($\mathbf{x} \in \mathbb{R}^D$):

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

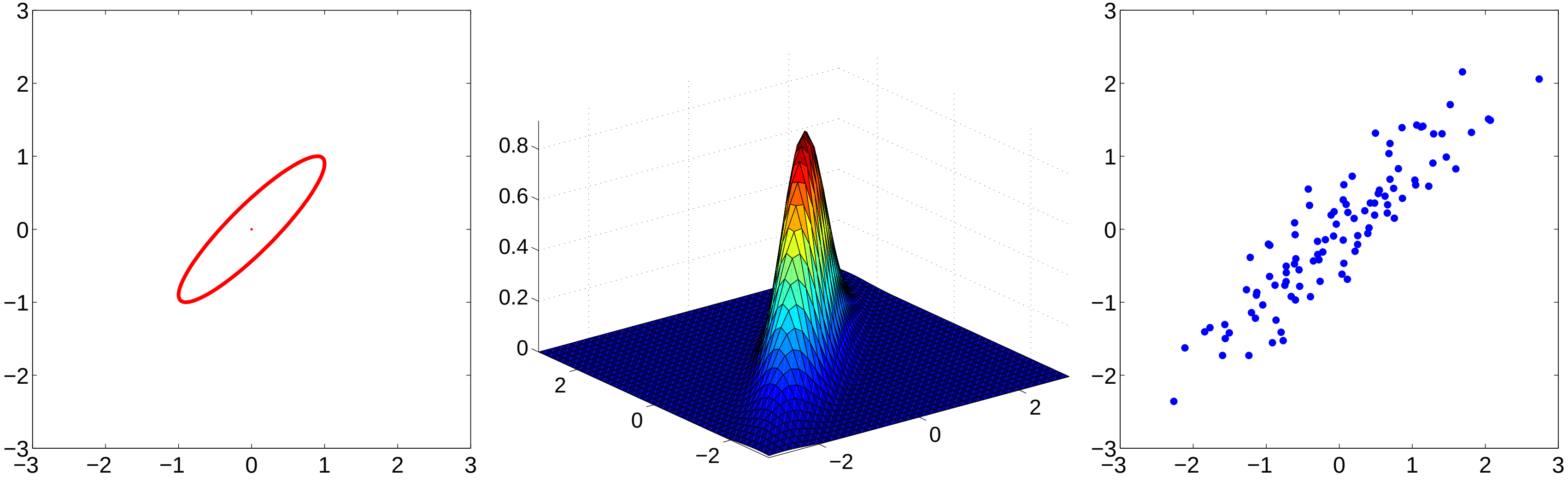
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



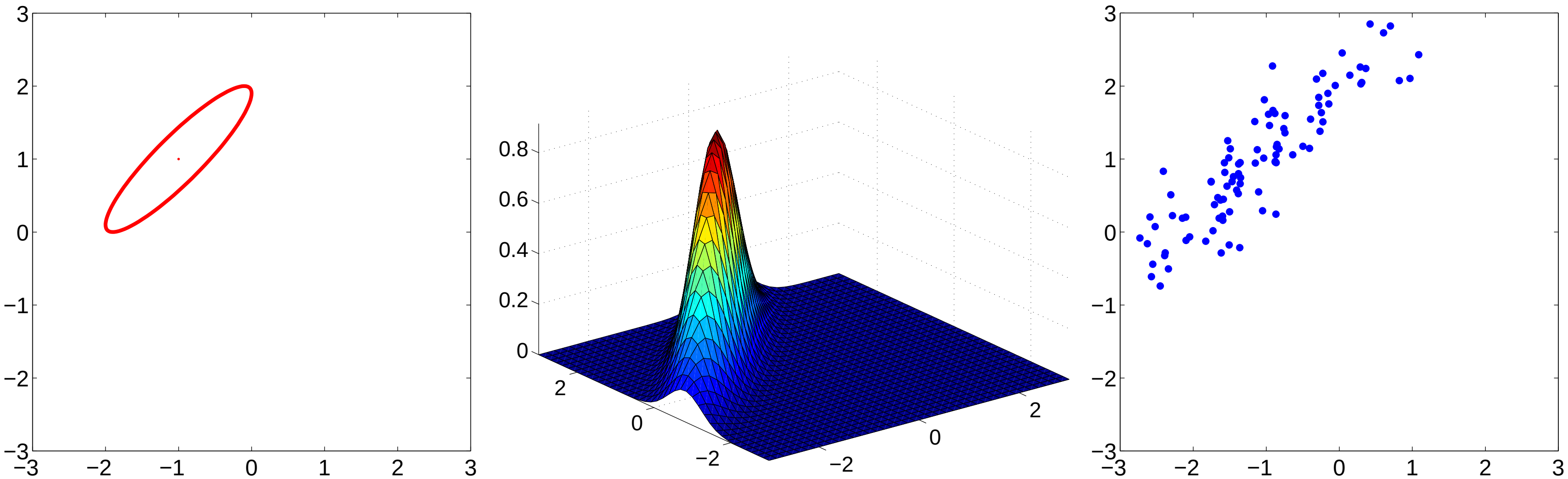
This model has parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \Sigma\}$ which model the mean and covariance matrix of the data.

The multivariate Gaussian density

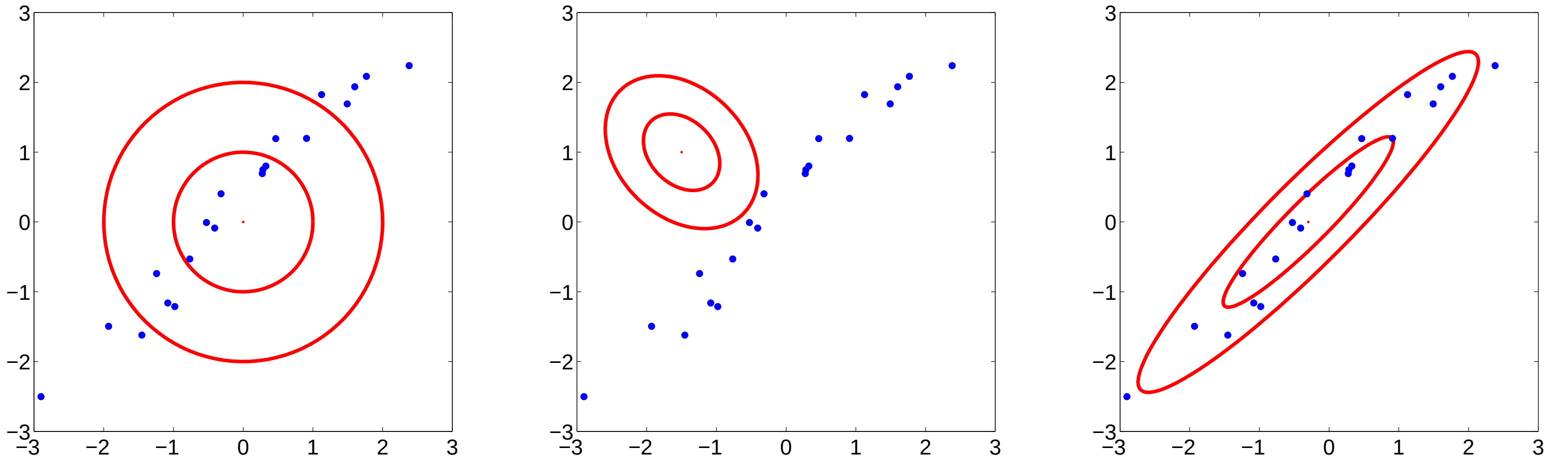
$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$



$$\boldsymbol{\mu} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$



Fitting a model to data



Assume the data were generated independently from the model.
We can measure the likelihood of the model:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\theta})$$

Clearly, the third model is a better fit to the data than the others:

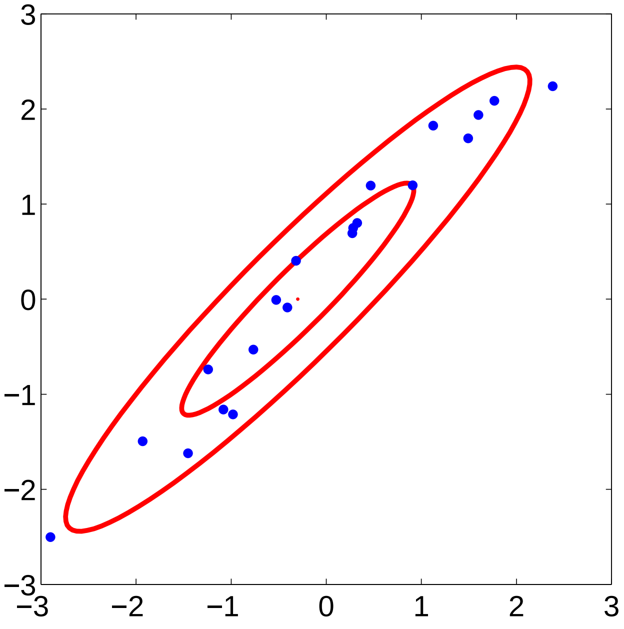
$$\begin{aligned}\log p(\mathcal{D}|\boldsymbol{\theta}_1) &= -55.38 \\ \log p(\mathcal{D}|\boldsymbol{\theta}_2) &= -238.29 \\ \log p(\mathcal{D}|\boldsymbol{\theta}_3) &= -22.14\end{aligned}$$

The likelihood function

Data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, the likelihood: $p(\mathcal{D}|\boldsymbol{\mu}, \Sigma) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}, \Sigma)$ is a function of the model parameters

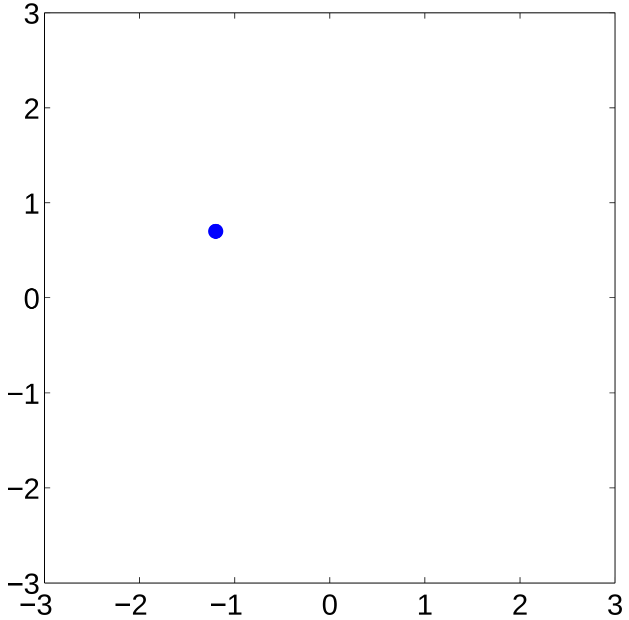
The maximum likelihood (ML) procedure finds parameters $\boldsymbol{\theta}_{\mathrm{ML}} = \{\boldsymbol{\mu}, \Sigma\}$ such that:

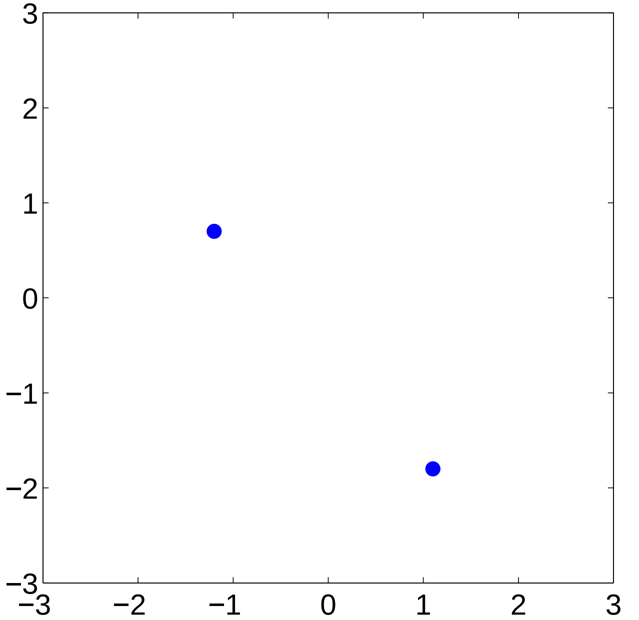
$$\boldsymbol{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, p(\mathcal{D}|\boldsymbol{\theta})$$



Two *very* simple data sets

What are the maximum likelihood estimates of θ for these data sets?





Does this make sense?

Bayesian Learning

Apply the basic rules of probability to learning from data.
Use probability distributions to represent uncertainty.

Data set: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
Model parameters: $\boldsymbol{\theta}$

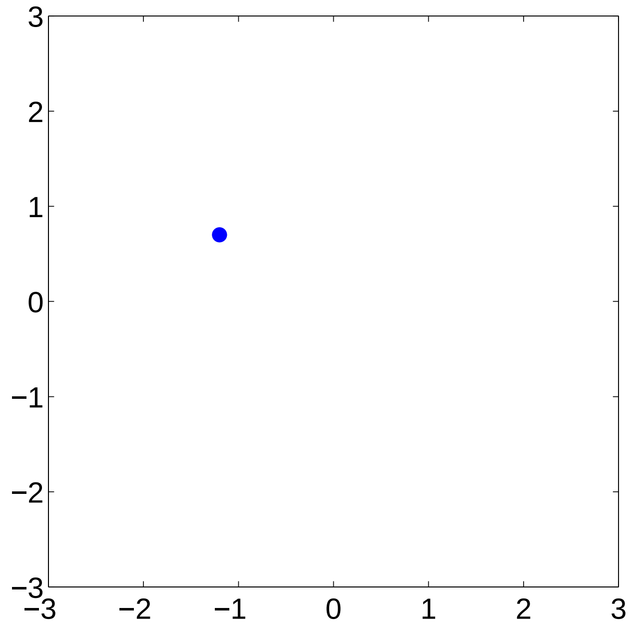
Prior probabilities of model parameters: $P(\boldsymbol{\theta})$
Model of data given parameters (likelihood model): $P(\mathbf{x}|\boldsymbol{\theta})$

If the data are independently and identically distributed then:

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\mathbf{x}_n|\boldsymbol{\theta})$$

Posterior probability of model parameters:

$$P(\boldsymbol{\theta}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathcal{D})}$$



Basic Rules of Probability

Let X be a random variable taking values x in some set $\mathcal{X}$.

Probabilities are non-negative $P(X = x) \geq 0 \; \forall x$.

Probabilities normalise: $\sum_{x \in \mathcal{X}} P(X = x) = 1$ for distributions if x is a discrete variable and $\int_{-\infty}^{+\infty} p(x) dx = 1$ for probability densities over continuous variables

The joint probability of $X = x$ and $Y = y$ is: $P(X = x, Y = y)$.

The marginal probability of $X = x$ is: $P(X = x) = \sum_y P(X = x, y)$, assuming y is discrete. I will generally write $P(x)$ to mean $P(X = x)$.

The conditional probability of x given y is: $P(x|y) = P(x, y)/P(y)$

Bayes Rule:

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y) \quad \Rightarrow \quad \boxed{P(y|x) = \frac{P(x|y)P(y)}{P(x)}}$$

Basic Rules of Probability and Bayesian Learning

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

Learning:

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}|\theta, m)$ likelihood of parameters θ in model m

$P(\theta|m)$ prior probability of θ

$P(\theta|\mathcal{D}, m)$ posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$