Machine Learning: Mathematical Background

Probability

Dariush Hosseini

dariush.hosseini@ucl.ac.uk
Department of Computer Science
University College London

Lecture Overview

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations $\longrightarrow$ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data).
    To this end we are often interested in studying the continuous change of such functions $\longrightarrow$ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically $\longrightarrow$ **Probability**
  - Drawing conclusions based on the analysis of data $\longrightarrow$ **Statistics**

Maths & Machine Learning

- Much of machine learning is concerned with:
  - Solving systems of linear equations $\longrightarrow$ **Linear Algebra**
  - Minimising cost functions (a scalar function of several variables that typically measures how poorly our model fits the data).
    To this end we are often interested in studying the continuous change of such functions $\longrightarrow$ **(Differential) Calculus**
  - Characterising uncertainty in our learning environments stochastically $\longrightarrow$ **Probability**
  - Drawing conclusions based on the analysis of data $\longrightarrow$ **Statistics**

Probabilistic Machine Learning

- The probabilistic setting is common in machine learning
- We will encounter this setting many times, so we should take the time to examine what learning looks like within it
- Typically (although not always) in this setting we suppose that the data is drawn from a **probability distribution** whose parameterisation we seek to learn
- We use data to learn (**infer**) a **point estimate** of the parameter in question

Learning Outcomes for Today's Lecture

- By the end of this lecture you should be familiar with some fundamental objects in and results of **Probability theory**
- For the most part we will concentrate on the statement of results which will be of use in the main body of this module
- However we will not be so concerned with the proof of these results

Lecture Overview

Introduction

- Probability is the measure of the likelihood, in some sense, that an event will occur
- The probability of the occurrence of an event is a number between 0 and 1
- The higher the probability of an event, the more likely that the event will occur

Setting

- Let us assume the existence of a **probability space**, $(\Omega, \mathcal{F}, \mathbb{P})$, consisting of:
  - A **sample space**, Ω, which is a fixed set of possible outcomes (e.g. the results of a coin toss or a dice roll)
  - An **event space**, $\mathcal{F}$, which is a set of **events**, where an event is a subset of Ω
    - The complement of an event A is another event $A^C = \Omega \backslash A$
  - A **probability measure**, $\mathbb{P}$ which is some mapping from $\mathcal{F}$ which satisfies the **probability axioms**

Probability Axioms

1. The probability of an event, E, is a non-negative real number:

$$\mathbb{P}(E) \in \mathbb{R}, \qquad \mathbb{P}(E) \geqslant 0, \qquad \forall\, E \in \mathcal{F}$$

2. $\mathbb{P}(\Omega) = 1$

3. For any countable collection of disjoint events, $\{E_i\}_{i=1}^{n}$:

$$\mathbb{P}\left(\bigcup_i E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i)$$

Consequences

1. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$
2. If A and B are events, and $B \subseteq A$, then $\mathbb{P}(B) < \mathbb{P}(A)$
3. $0 = \mathbb{P}(\varnothing) \leqslant \mathbb{P}(A) \leqslant \mathbb{P}(\Omega) = 1$
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(A) - \mathbb{P}(A \cap B)$
5. **Union Bound:**
   If $\{E_i\}_{i=1}^n$ is a countable set of events, disjoint or not, then:

$$\mathbb{P}\left(\bigcup_i E_i\right) \leqslant \sum_{i=1}^{n} \mathbb{P}(E_i)$$

Conditional Probability & Bayes' Rule

- The **conditional probability** of an event A, given that an event B has occurred is written as:

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

- From which **Bayes' Rule** follows:

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

  - $\mathbb{P}(A)$ is often referred to as the **prior**
  - $\mathbb{P}(A|B)$ is often referred to as the **posterior**
  - $\mathbb{P}(B|A)$ is often referred to as the **likelihood**
  - $\mathbb{P}(B)$ is often referred to as the **evidence**

Lecture Overview

Random Variables

- Informally, a **random variable** is a variable where possible values are outcomes of a random phenomenon
- Formally, a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $\mathcal{X} : \Omega \to \mathbb{R}$
- The range of $\mathcal{X}$ is given by:

$$\mathcal{X}(\Omega) = \{\mathcal{X}(\omega) | \omega \in \Omega\}$$

Random Variables

- The probability that an outcome or **sample** associated with $\mathcal{X}$ assumes a value in a set $S \subseteq \mathbb{R}$ is written as:

$$\mathbb{P}(\mathcal{X} \in S) = \mathbb{P}(\{\omega \in \Omega | \mathcal{X}(\omega) \in S\})$$

  - e.g. if $\mathcal{X}$ is the number of heads in 2 tosses of a coin then:

$$\Omega = \{HH, TT, HT, TH\} = \{2, 0, 1, 1\}$$

    The event, $S = 1$, then has probability:

$$\mathbb{P}(\mathcal{X} = 1) = \mathbb{P}(\{HT, TH\}) = \frac{1}{2}$$

Cumulative Distribution Function

- The **cumulative distribution function** (cdf) gives the probability that the outcome of a random variable is at most a particular value, x:

$$F_{\mathcal{X}}(x) = \mathbb{P}(\mathcal{X} \leqslant x)$$

Probability Distribution

- The complete specification of probability for the different values that samples taken from $\mathcal{X}$ can assume is called the **probability distribution** of $\mathcal{X}$, $\mathcal{D}$, and is characterised by some mapping, $p_{\mathcal{X}}$
- We say that $\mathcal{X}$ is a random variable with outcomes, x, sampled according to (or **drawn from**) $\mathcal{D}$:

$$x \sim \mathcal{D}$$

  And sometimes we write $\mathbb{P}_{\mathcal{D}}$ instead of $\mathbb{P}$ to emphasise this
- The precise characterisation of this function, $p_{\mathcal{X}}$, depends on whether $\mathcal{X}$ is **discrete** or **continuous**

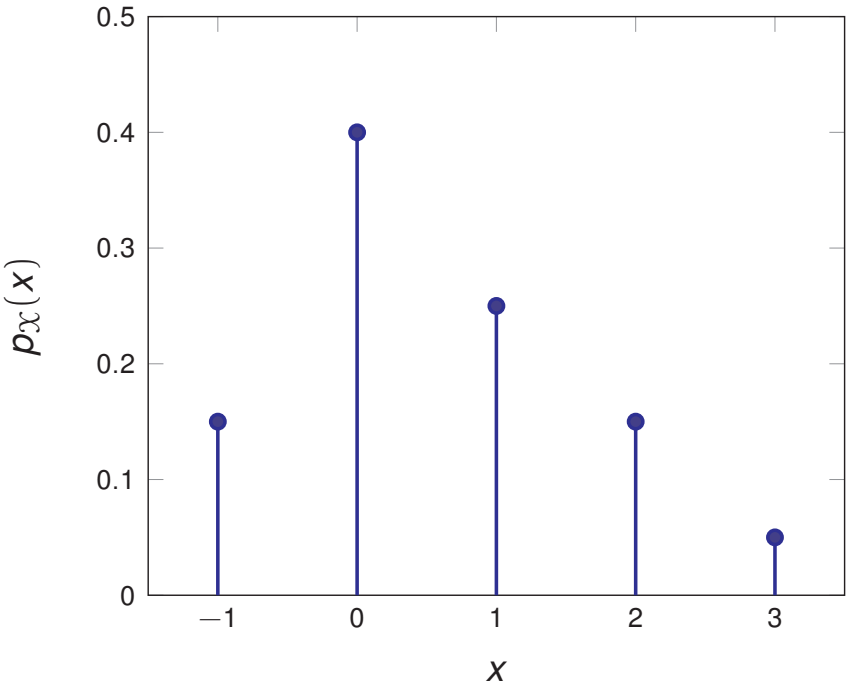
Discrete Random Variable

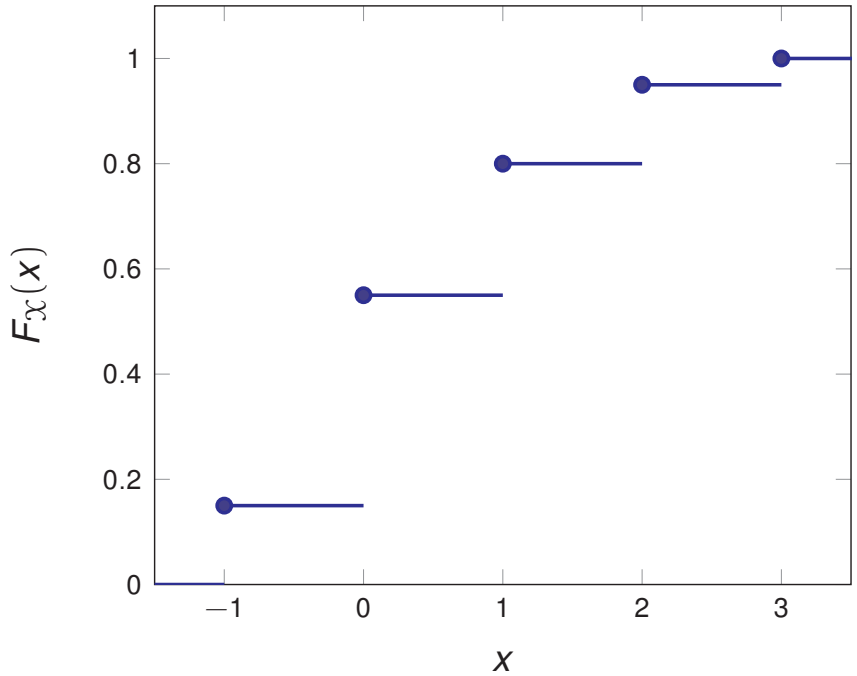
- A **discrete random variable** is a random variable that has a countable range
- It is specified by its **probability mass function** (pmf), $p_X : \mathcal{X}(\Omega) \to [0, 1]$, so:

$$\mathbb{P}_{\mathcal{D}}(\mathcal{X} = x) = p_X(x)$$

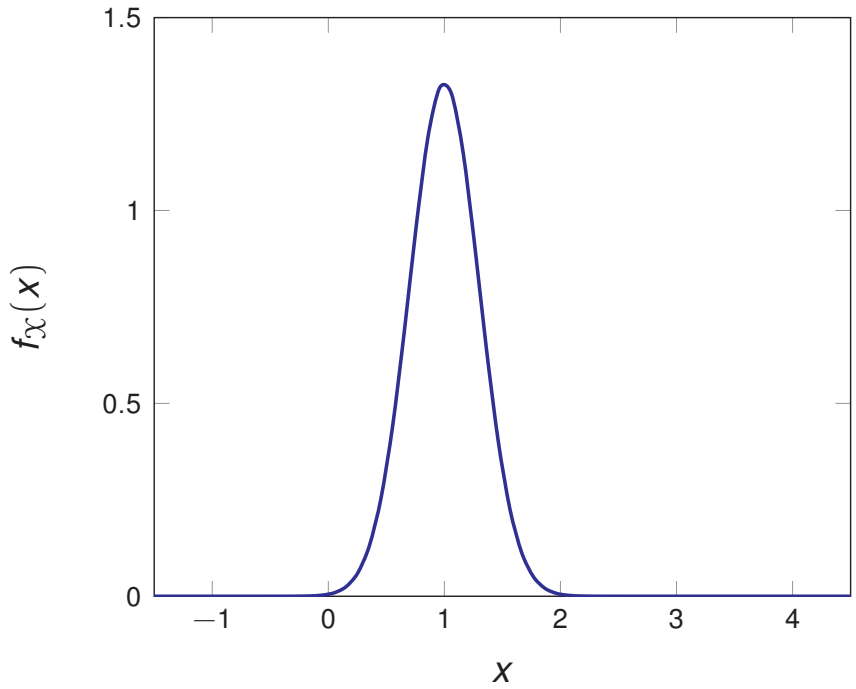
$$\sum_{x \in \mathcal{X}(\Omega)} p_X(x) = 1$$

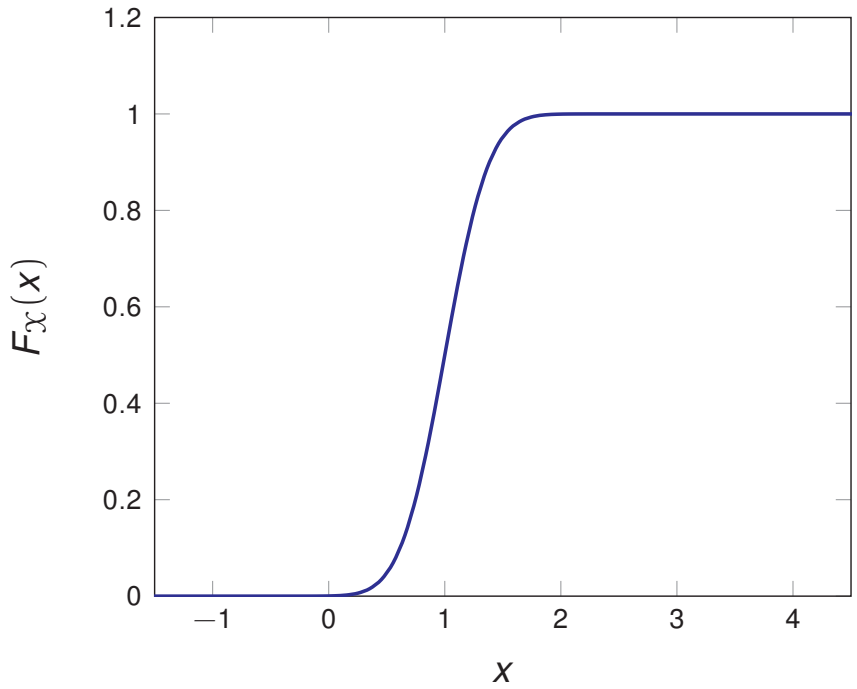
Discrete Random Variable: Example





Continuous Random Variable: Example





Continuous Random Variable

- A **continuous random variable** is a random variable that has an uncountable range
- It is specified by its **probability density function** (pdf), $p_{\mathcal{X}} : \mathbb{R} \to [0, \infty)$, (although I often write, $f_{\mathcal{X}} : \mathbb{R} \to [0, \infty)$), so:

$$\mathbb{P}_{\mathcal{D}}(a \leqslant \mathcal{X} \leqslant b) = \int_a^b f_{\mathcal{X}}(x)dx$$

$$\int_{-\infty}^{\infty} f_{\mathcal{X}}(x)dx = 1$$

Joint Distributions

- A **joint** or **multivariate distribution** is the probability distribution, $\mathcal{D}$, over some combination of m random variables, $\{\mathcal{X}_i\}_{i=1}^m$, with outcomes $\{x_i \sim \mathcal{D}_i\}_{i=1}^m$

- This probability distribution is characterised by some mapping, $p_{\mathcal{X}_1,\ldots,\mathcal{X}_m}$

- Generally, we express a collection of such random variables as a **random vector**, $\boldsymbol{\mathcal{X}}$, with outcomes, $\mathbf{x}$ so that:

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} \mathcal{X}_1 \\ \mathcal{X}_2 \\ \vdots \\ \mathcal{X}_m \end{bmatrix}; \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}$$

Independence

- We say that 2 random variables, $\mathcal{X}$ and $\mathcal{Y}$ are **independent** if their joint distribution factors:

$$p_{\mathcal{X},\mathcal{Y}}(\cdot,\cdot) = p_{\mathcal{X}}(\cdot)p_{\mathcal{Y}}(\cdot)$$

- In general, if $\mathcal{X}_1, ..., \mathcal{X}_m$ are independent, then:

$$p_{\mathcal{X}_1,...,\mathcal{X}_m}(\cdot,...,\cdot) = \prod_{i=1}^{m} p_{\mathcal{X}_i}(\cdot)$$

Independence

- A collection of random variables is **independent and identically distributed** (**i.i.d.**) if each random variable has the same probability distribution as the others and all are mutually independent

- If a collection of random variables, $\mathcal{X}_1, ..., \mathcal{X}_m$, are i.i.d., then I often write the following short-hand as notation for their joint distribution:

$$p_{\mathcal{X}_1,...,\mathcal{X}_m}(\cdot, ..., \cdot) = p_{\mathcal{X}}(\cdot, ..., \cdot)$$

Marginal Distribution

- If we have a joint distribution over a set of random variables then we can obtain a **marginal distribution** for a subset of them by summing out the redundant variable:
  - For two discrete random variables, $\mathcal{X}$, $\mathcal{Y}$:

$$p_{\mathcal{X}}(x) = \sum_{y \in \mathcal{Y}(\Omega)} p_{\mathcal{X},\mathcal{Y}}(x, y)$$

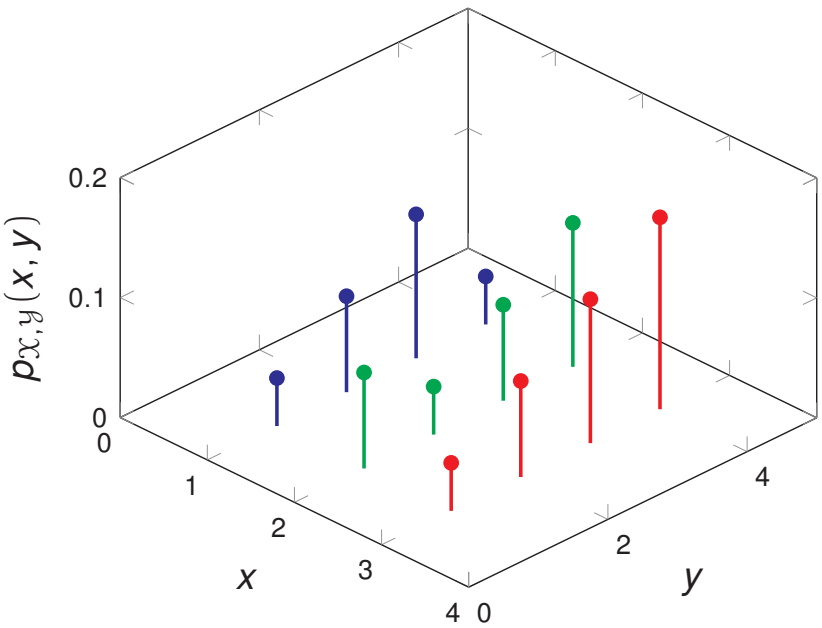
  - For two continuous random variables, $\mathcal{X}$, $\mathcal{Y}$:

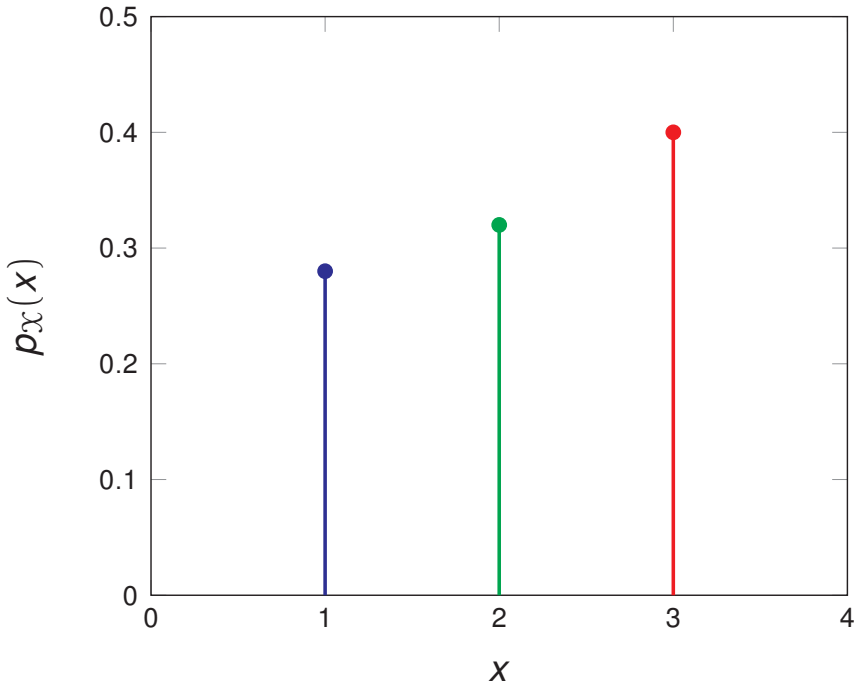
$$f_{\mathcal{X}}(x) = \int_{\mathcal{Y}(\Omega)} f_{\mathcal{X},\mathcal{Y}}(x, y) dy$$

Marginal Distribution: Discrete Example

$p_{X,Y}(x,y)$	$x=1$	$x=2$	$x=3$	$p_Y(y)$
$y=1$	0.04	0.08	0.04	0.16
$y=2$	0.08	0.04	0.08	0.20
$y=3$	0.12	0.08	0.12	0.32
$y=4$	0.04	0.12	0.16	0.32
$p_X(x)$	0.28	0.32	0.40	1.00

Marginal Distribution: Discrete Example





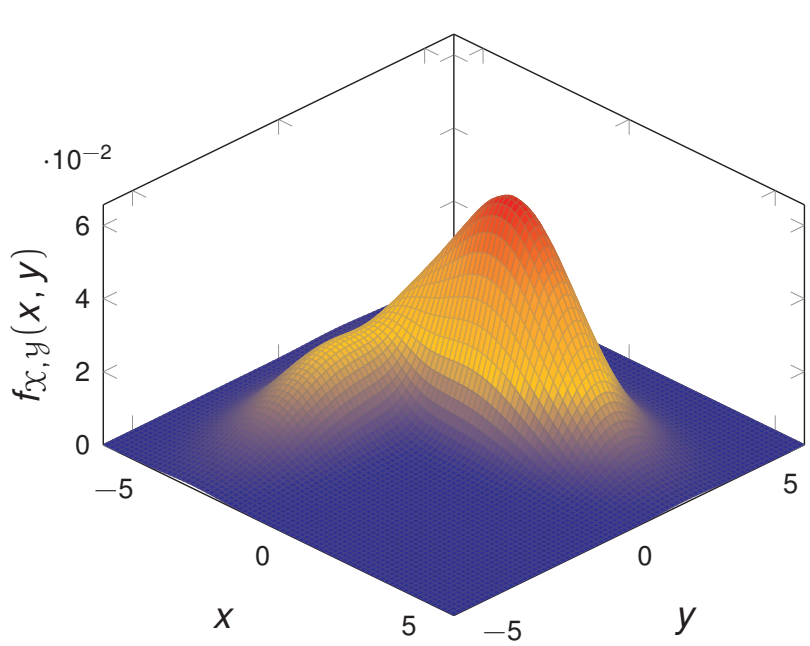
Marginal Distribution: Continuous Example

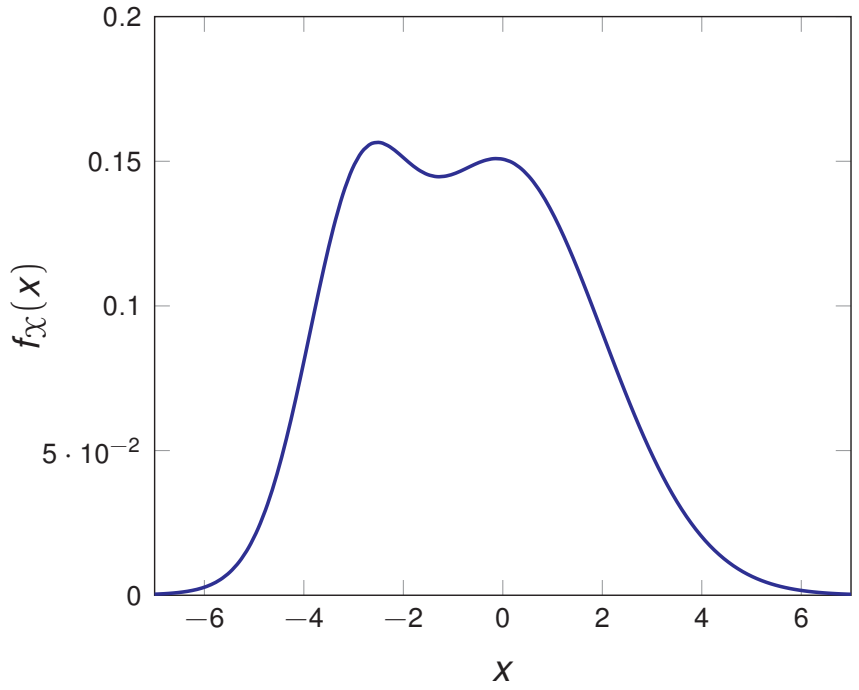
$$f_{\mathcal{X},\mathcal{Y}}(x,y) = \frac{\alpha}{2\pi\sigma_{1x}\sigma_{1y}} \exp -\frac{1}{2}\left(\frac{(x-\mu_{1x})^2}{\sigma_{1x}^2} + \frac{(y-\mu_{1y})^2}{\sigma_{1y}^2}\right) + \frac{1-\alpha}{2\pi\sigma_{2x}\sigma_{2y}} \exp -\frac{1}{2}\left(\frac{(x-\mu_{2x})^2}{\sigma_{2x}^2} + \frac{(y-\mu_{2y})^2}{\sigma_{2y}^2}\right)$$

Where:
$\alpha = 0.75$,
$[\mu_{1x}, \mu_{1y}] = [0, 2]$, $[\sigma_{1x}, \sigma_{1y}] = [2, 1]$,
$[\mu_{2x}, \mu_{2y}] = [-3, -1]$, $[\sigma_{2x}, \sigma_{2y}] = [1, 2]$

$$\begin{aligned} f_{\mathcal{X}}(x) &= \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x,y)dy \\ &= \frac{\alpha}{\sqrt{2\pi}\sigma_{1x}} \exp -\frac{1}{2}\left(\frac{(x-\mu_{1x})^2}{\sigma_{1x}^2}\right) + \frac{1-\alpha}{\sqrt{2\pi}\sigma_{2x}} \exp -\frac{1}{2}\left(\frac{(x-\mu_{2x})^2}{\sigma_{2x}^2}\right) \end{aligned}$$

Marginal Distribution: Continuous Example





Conditional Distributions

- If we have a joint distribution over some combination of random variables, $\mathcal{X}, \mathcal{Y}$, then we can obtain the distribution of $\mathcal{X}$ **conditional** on an outcome of $\mathcal{Y}$ as follows:
  - For two discrete random variables, $\mathcal{X}$, $\mathcal{Y}$:

$$p_{\mathcal{X}}(x|y) = \frac{p_{\mathcal{X},\mathcal{Y}}(x,y)}{\sum_{x \in \mathcal{X}(\Omega)} p_{\mathcal{X},\mathcal{Y}}(x,y)}$$

  - For two continuous random variables, $\mathcal{X}$, $\mathcal{Y}$:

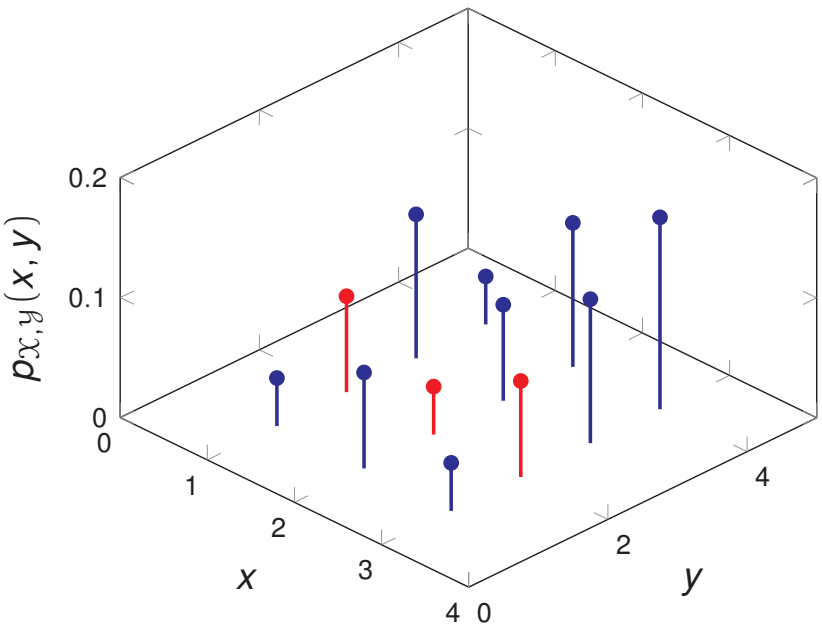
$$f_{\mathcal{X}}(x|y) = \frac{f_{\mathcal{X},\mathcal{Y}}(x,y)}{\int_{\mathcal{X}(\Omega)} f_{\mathcal{X},\mathcal{Y}}(x,y)dx}$$

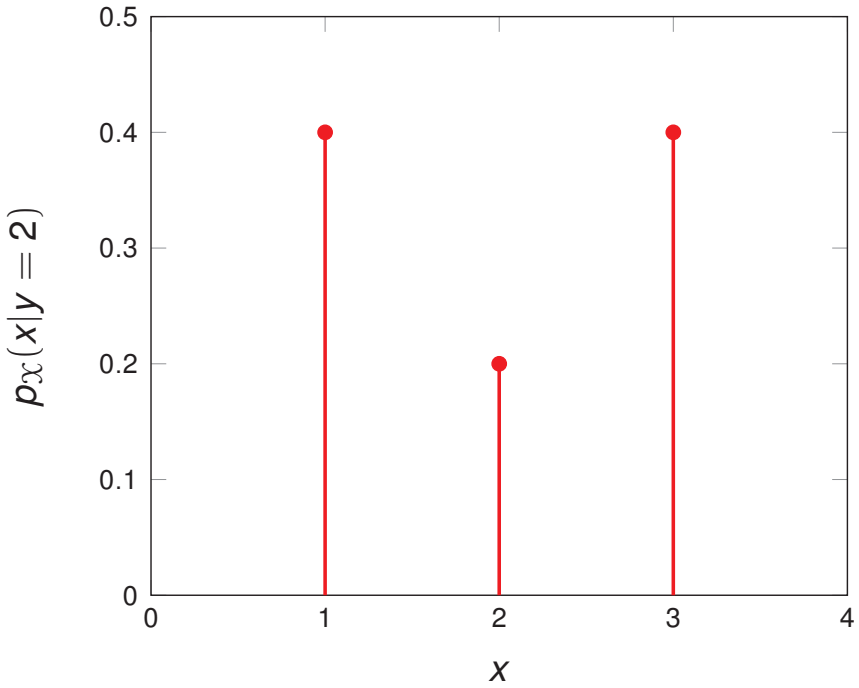
Conditional Distribution: Discrete Example

$p_{X,Y}(x,y)$	$x=1$	$x=2$	$x=3$	$p_Y(y)$
$y=1$	0.04	0.08	0.04	0.16
$y=2$	0.08	0.04	0.08	0.20
$y=3$	0.12	0.08	0.12	0.32
$y=4$	0.04	0.12	0.16	0.32
$p_X(x)$	0.28	0.32	0.40	1.00

	$x=1$	$x=2$	$x=3$	$p_Y(y)$
$p_X(x\|y=2)$	0.40	0.20	0.40	1.00

Conditional Distribution: Discrete Example





Conditional Distribution: Continuous Example

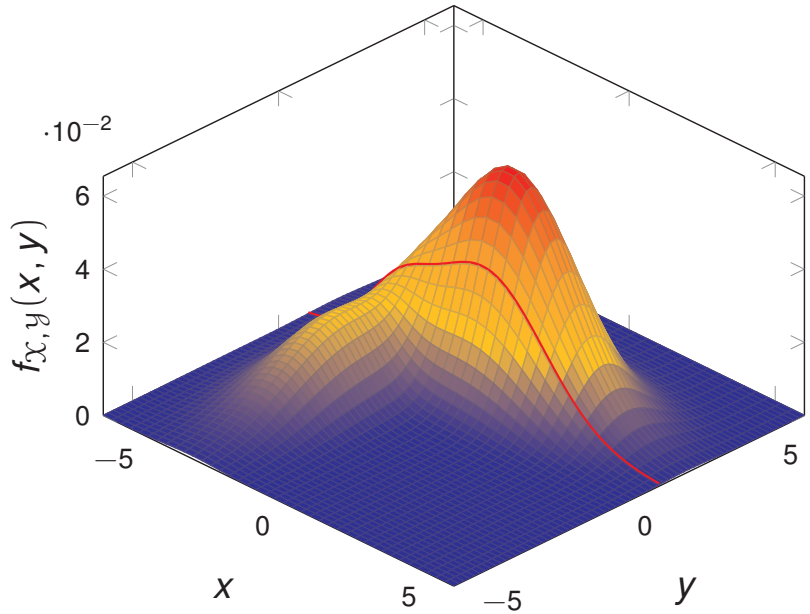
$$f_{\mathcal{X}}(x|y=1) = \frac{f_{\mathcal{X},\mathcal{Y}}(x, y=1)}{f_{\mathcal{Y}}(y=1)}$$

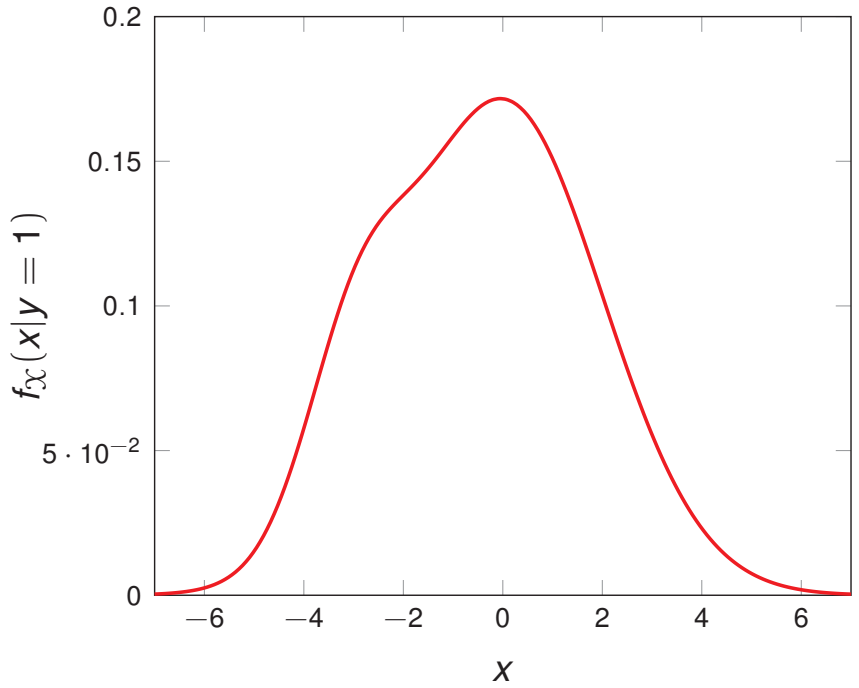
Where:

$$f_{\mathcal{X},\mathcal{Y}}(x, y=1) = \frac{\alpha}{2\pi\sigma_{1x}\sigma_{1y}} \exp -\frac{1}{2}\left(\frac{(x-\mu_{1x})^2}{\sigma_{1x}^2} + \frac{(1-\mu_{1y})^2}{\sigma_{1y}^2}\right) + \frac{1-\alpha}{2\pi\sigma_{2x}\sigma_{2y}} \exp -\frac{1}{2}\left(\frac{(x-\mu_{2x})^2}{\sigma_{2x}^2} + \frac{(1-\mu_{2y})^2}{\sigma_{2y}^2}\right)$$

$$\begin{aligned} f_{\mathcal{Y}}(y=1) &= \int_{-\infty}^{\infty} f_{\mathcal{X},\mathcal{Y}}(x, y=1)dx \\ &= \frac{\alpha}{\sqrt{2\pi}\sigma_{1y}} \exp -\frac{1}{2}\left(\frac{(1-\mu_{1y})^2}{\sigma_{1y}^2}\right) + \frac{1-\alpha}{\sqrt{2\pi}\sigma_{2y}} \exp -\frac{1}{2}\left(\frac{(1-\mu_{2y})^2}{\sigma_{2y}^2}\right) \\ &= 0.2117 \end{aligned}$$

Conditional Distribution: Continuous Example





Lecture Overview

Expected Value

- The **expected value** or **mean** of some random variable, $\mathcal{X}$, with outcomes, $x \sim \mathcal{D}$, is defined as:
  - For discrete random variables:

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = \sum_{x \in \mathcal{X}(\Omega)} x p_{\mathcal{X}}(x)$$

  - For continuous random variables:

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = \int_{-\infty}^{\infty} x f_{\mathcal{X}}(x) dx$$

Expected Value: Functions of a Random Variable

- For some function, g, and some random variable, $\mathcal{X}$, with outcomes, $x \sim \mathcal{D}$, the **expected value** of the random variable, $g(\mathcal{X})$, is:
  - For discrete random variables:

$$\mathbb{E}_{\mathcal{D}}[g(\mathcal{X})] = \sum_{x \in \mathcal{X}(\Omega)} g(x) p_{\mathcal{X}}(x)$$

  - For continuous random variables:

$$\mathbb{E}_{\mathcal{D}}[g(\mathcal{X})] = \int_{-\infty}^{\infty} g(x) f_{\mathcal{X}}(x) dx$$

Expected Value: Multivariate

- The expected value of some random vector, $\mathfrak{X} = [\mathfrak{X}_1, \mathfrak{X}_2, \ldots, \mathfrak{X}_m]^T$, with outcomes, $\mathbf{x} \sim \mathcal{D}$, where $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ and $\{x_i \sim \mathcal{D}_i\}_{i=1}^m$ is:

$$\mathbb{E}_{\mathcal{D}}[\mathfrak{X}] = \begin{bmatrix} \mathbb{E}_{\mathcal{D}_1}[\mathfrak{X}_1] \\ \mathbb{E}_{\mathcal{D}_2}[\mathfrak{X}_2] \\ \vdots \\ \mathbb{E}_{\mathcal{D}_m}[\mathfrak{X}_m] \end{bmatrix}$$

Expected Value: Properties

- We can prove that in general, for a set of random variables, $\{\mathcal{X}_i\}_{i=1}^m$, with outcomes, $\{x_i \sim \mathcal{D}_i\}_{i=1}^m$, drawn from a joint distribution $\mathcal{D}$ (even if $\mathcal{X}_i$ are not independent):

$$\mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{m} \alpha_i \mathcal{X}_i + \beta\right] = \sum_{i=1}^{m} \alpha_i \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{X}_i\right] + \beta$$

  where $\alpha_i, \beta \in \mathbb{R}$

- And in particular for independent random variables:

$$\mathbb{E}_{\mathcal{D}}\left[\prod_{i=1}^{m} \mathcal{X}_i\right] = \prod_{i=1}^{m} \mathbb{E}_{\mathcal{D}_i}\left[\mathcal{X}_i\right]$$

Variance

- The **variance** of some random variable, $\mathcal{X}$, with outcomes, $x \sim \mathcal{D}$, is defined as:

$$\begin{aligned}\mathrm{Var}_{\mathcal{D}}[\mathcal{X}] &= \mathbb{E}_{\mathcal{D}}\left[(\mathcal{X} - \mathbb{E}_{\mathcal{D}}[\mathcal{X}])^2\right] \\ &= \mathbb{E}_{\mathcal{D}}[\mathcal{X}^2] - (\mathbb{E}_{\mathcal{D}}[\mathcal{X}])^2\end{aligned}$$

- The **standard deviation** is defined as $\sqrt{\mathrm{Var}_{\mathcal{D}}[\mathcal{X}]}$

Variance: Properties

- We can prove that in general:

$$\text{Var}_{\mathcal{D}}\left[\alpha\mathcal{X} + \beta\right] = \alpha^2\text{Var}_{\mathcal{D}}\left[\mathcal{X}\right]$$

  where $\alpha, \beta \in \mathbb{R}$

- And in particular for **uncorrelated** random variables, $\{\mathcal{X}_i\}_{i=1}^m$, with outcomes $\{x_i \sim \mathcal{D}_i\}_{i=1}^m$, drawn from a joint distribution $\mathcal{D}$:

$$\text{Var}_{\mathcal{D}}\left[\mathcal{X}_1 + ... + \mathcal{X}_m\right] = \text{Var}_{\mathcal{D}_1}\left[\mathcal{X}_1\right] + ... + \text{Var}_{\mathcal{D}_m}\left[\mathcal{X}_m\right]$$

Covariance: Univariate

- The **covariance** is a measure of the linear relationship between two random variables, $\mathcal{X}_1$ (with outcomes $x_1 \sim \mathcal{D}_1$) and $\mathcal{X}_2$ (with outcomes $x_2 \sim \mathcal{D}_2$), drawn from a joint distribution, $\mathcal{D}$ and is defined as:

$$\text{Cov}[\mathcal{X}_1, \mathcal{X}_2] = \mathbb{E}_{\mathcal{D}}\left[(\mathcal{X}_1 - \mathbb{E}_{\mathcal{D}_1}[\mathcal{X}_1])(\mathcal{X}_2 - \mathbb{E}_{\mathcal{D}_2}[\mathcal{X}_2])\right]$$

- We can prove that:

$$\text{Cov}[\mathcal{X}_1, \mathcal{X}_2] = \mathbb{E}_{\mathcal{D}}\left[\mathcal{X}_1\mathcal{X}_2\right] - \mathbb{E}_{\mathcal{D}_1}\left[\mathcal{X}_1\right]\mathbb{E}_{\mathcal{D}_2}\left[\mathcal{X}_2\right]$$

- And for random variables $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$:

$$\text{Cov}[\alpha\mathcal{X} + \beta\mathcal{Y}, \mathcal{Z}] = \alpha\text{Cov}[\mathcal{X}, \mathcal{Z}] + \beta\text{Cov}[\mathcal{Y}, \mathcal{Z}]$$

Correlation: Univariate

- The **correlation** is the normalised covariance and always lies between -1 and 1:

$$\rho(\mathcal{X}_1, \mathcal{X}_2) = \frac{\text{Cov}[\mathcal{X}_1, \mathcal{X}_2]}{\sqrt{\text{Var}_{\mathcal{D}_1}[\mathcal{X}_1]\text{Var}_{\mathcal{D}_2}[\mathcal{X}_2]}}$$

- Two variables are said to be **uncorrelated** if:

$$\text{Cov}[\mathcal{X}_1, \mathcal{X}_2] = \rho(\mathcal{X}_1, \mathcal{X}_2) = 0$$

- If two variables are independent $\implies$ they are uncorrelated
- If two variables are uncorrelated $\not\implies$ they are independent in general

Covariance: Multivariate

- The covariance of some random vector, $\mathfrak{X} = [\mathfrak{X}_1, \mathfrak{X}_2, \ldots, \mathfrak{X}_m]^T$, with outcomes, $\mathbf{x} \sim \mathcal{D}$, where $\mathbf{x} = [x_1, x_2, \ldots, x_m]^T$ and $\{x_i \sim \mathcal{D}_i\}_{i=1}^m$ is characterised by the **covariance matrix**:

$$\boldsymbol{\Sigma} = \mathbb{E}_{\mathcal{D}}\left[(\mathfrak{X} - \mathbb{E}_{\mathcal{D}}[\mathfrak{X}])(\mathfrak{X} - \mathbb{E}_{\mathcal{D}}[\mathfrak{X}])^T\right]$$

$$= \begin{bmatrix} \mathrm{Var}[\mathfrak{X}_1] & \mathrm{Cov}[\mathfrak{X}_1, \mathfrak{X}_2] & \ldots & \mathrm{Cov}[\mathfrak{X}_1, \mathfrak{X}_m] \\ \mathrm{Cov}[\mathfrak{X}_2, \mathfrak{X}_1] & \mathrm{Var}[\mathfrak{X}_2] & \ldots & \mathrm{Cov}[\mathfrak{X}_2, \mathfrak{X}_m] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[\mathfrak{X}_m, \mathfrak{X}_1] & \mathrm{Cov}[\mathfrak{X}_m, \mathfrak{X}_2] & \ldots & \mathrm{Var}[\mathfrak{X}_m] \end{bmatrix}$$

- The **precision matrix** is defined as $\boldsymbol{\Sigma}^{-1}$

Covariance: Multivariate

- Note that Σ is symmetric:

$$\Sigma^T = \Sigma$$

- And positive semidefinite:

$$\begin{aligned}
\mathbf{a}^T \Sigma \mathbf{a} &= \mathbf{a}^T \mathbb{E}_{\mathcal{D}} \left[(\mathbf{x} - \mathbb{E}_{\mathcal{D}}[\mathbf{x}])(\mathbf{x} - \mathbb{E}_{\mathcal{D}}[\mathbf{x}])^T \right] \mathbf{a} \\
&= \mathbb{E}_{\mathcal{D}} \left[\mathbf{a}^T (\mathbf{x} - \mathbb{E}_{\mathcal{D}}[\mathbf{x}])(\mathbf{x} - \mathbb{E}_{\mathcal{D}}[\mathbf{x}])^T \mathbf{a} \right] \\
&= \mathbb{E}_{\mathcal{D}} \left[\| (\mathbf{x} - \mathbb{E}_{\mathcal{D}}[\mathbf{x}])^T \mathbf{a} \|_2^2 \right] \geqslant 0
\end{aligned}$$

for all $\mathbf{a}$

Correlation: Multivariate

- Similarly, the multivariate correlation is characterised by the **correlation matrix**, a normalised version of the covariance matrix:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \frac{\mathrm{Cov}[x_1, x_2]}{\sqrt{\mathrm{Var}_{\mathcal{D}_1}[x_1]\mathrm{Var}_{\mathcal{D}_2}[x_2]}} & \cdots & \frac{\mathrm{Cov}[x_1, x_m]}{\sqrt{\mathrm{Var}_{\mathcal{D}_1}[x_1]\mathrm{Var}_{\mathcal{D}_m}[x_m]}} \\ \frac{\mathrm{Cov}[x_2, x_1]}{\sqrt{\mathrm{Var}_{\mathcal{D}_2}[x_2]\mathrm{Var}_{\mathcal{D}_1}[x_1]}} & 1 & \cdots & \frac{\mathrm{Cov}[x_2, x_m]}{\sqrt{\mathrm{Var}_{\mathcal{D}_2}[x_2]\mathrm{Var}_{\mathcal{D}_m}[x_m]}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathrm{Cov}[x_m, x_1]}{\sqrt{\mathrm{Var}_{\mathcal{D}_m}[x_m]\mathrm{Var}_{\mathcal{D}_1}[x_1]}} & \frac{\mathrm{Cov}[x_m, x_2]}{\sqrt{\mathrm{Var}_{\mathcal{D}_m}[x_m]\mathrm{Var}_{\mathcal{D}_2}[x_2]}} & \cdots & 1 \end{bmatrix}$$

Lecture Overview

Sample Statistics

- The earlier definitions for the expectation, variance and covariance are usually termed **population statistics**, since they refer to summary properties for the entire population of data outcomes

- We can define analagous quantities for more limited empirical samples of data outcomes, and we term these **sample statistics**

Sample Expectation

- Given some random vector, $\mathcal{X}$, with outcomes, $\mathbf{x} \in \mathbb{R}^m$, and a sample data set $\mathcal{S} = \{\mathbf{x}^{(i)}\}_{i=1}^n$, then the **sample expectation** or **sample mean** is the arithmetic average of the outcomes:

$$\overline{\mathbf{x}} = \mathbb{E}_{\mathcal{S}}[\mathcal{X}] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}^{(i)}$$

Sample Covariance

- Similarly the **sample covariance** is an $(m \times m)$ matrix, characterised as follows:

$$\begin{aligned}\mathbb{E}_S[\mathcal{X}] &= \mathbb{E}_S\left[(\mathcal{X} - \mathbb{E}_S[\mathcal{X}])(\mathcal{X} - \mathbb{E}_S[\mathcal{X}])^T\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}^{(i)} - \overline{\mathbf{x}})(\mathbf{x}^{(i)} - \overline{\mathbf{x}})^T\end{aligned}$$

- This matrix is symmetric and positive semidefinite

Lecture Overview

Bernoulli Distribution

- Here $\mathcal{X}$ is a discrete random variable, with outcomes, x, which can take one of two values: 0 or 1, with a probability that $x = 1$ being equal to θ

- This characterises $\mathcal{D}$ as a **Bernoulli Distribution**:

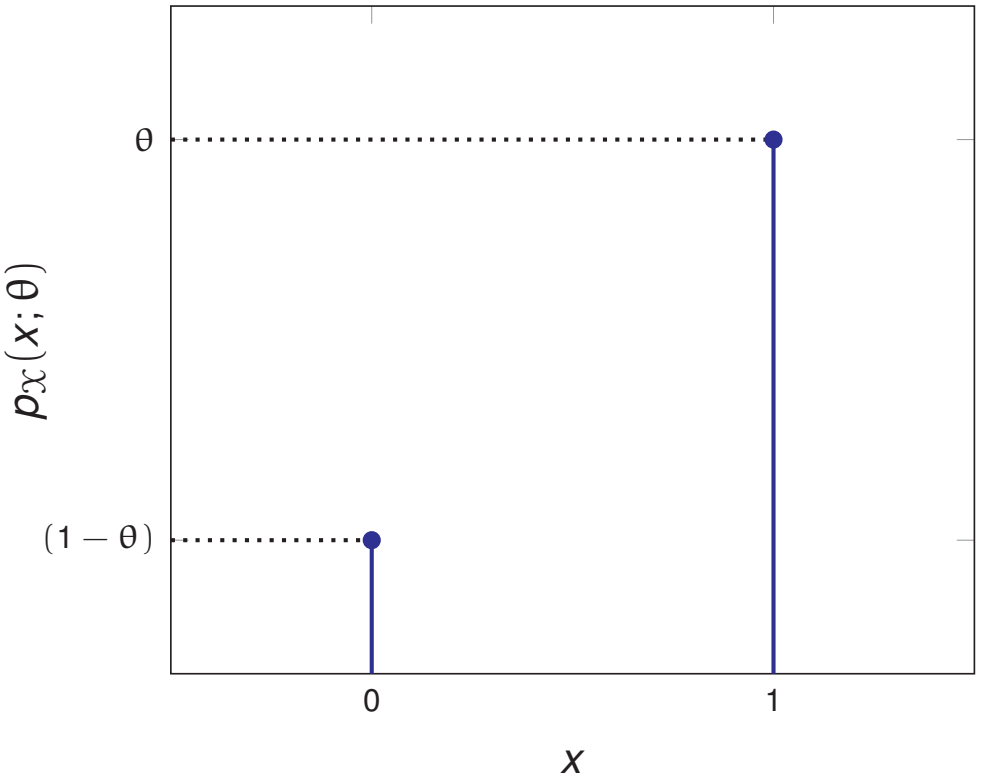
$$x \sim \text{Bern}(\theta)$$

$$p_{\mathcal{X}}(x; \theta) = \theta^x (1 - \theta)^{1-x}$$

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = \theta$$

$$\text{Var}_{\mathcal{D}}[\mathcal{X}] = \theta(1 - \theta)$$

Bernoulli Distribution



Binomial Distribution

- Here $\mathcal{X}$ is a discrete random variable, with outcomes, x, that denotes the number of successes, k, that we will achieve in n independent trials, where each trial is either a 'success' (=1) or a 'failure' (=0)

- $\mathcal{X}$ can thus be written as the **sum** of n **independent** Bernoulli trials $\{\mathcal{X}_i\}_{i=1}^n$, with outcomes $\{x_i \sim \text{Bern}(\theta)\}_{i=1}^n$:

$$\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2 + ... + \mathcal{X}_n$$

- This characterises the associated $\mathcal{D}$ as a **Binomial Distribution**:

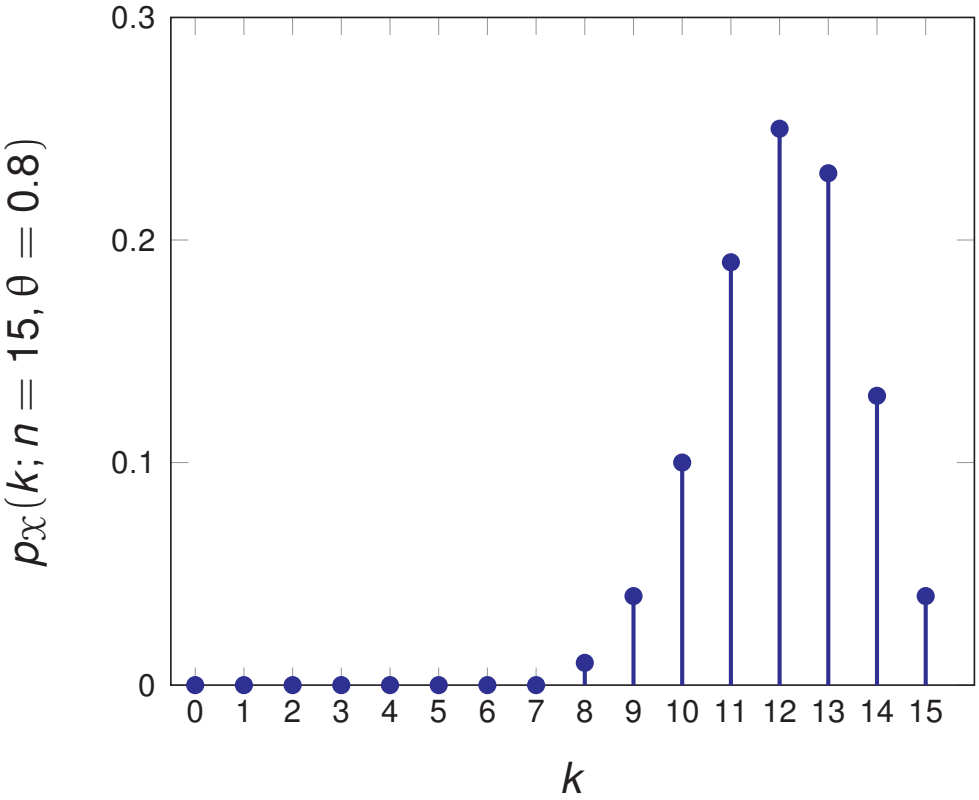
$$x \sim \text{Bin}(n, \theta)$$

$$p_{\mathcal{X}}(k; n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = n\theta$$

$$\text{Var}_{\mathcal{D}}[\mathcal{X}] = n\theta(1-\theta)$$

Binomial Distribution: Example



Gaussian Distribution

- Here $\mathcal{X}$ is a continuous random variable, taking values $x \in \mathbb{R}$, such that $\mathcal{D}$ follows a **Gaussian** or **Normal** distribution:

$$x \sim \mathcal{N}(\mu, \sigma^2) \qquad \text{where:} \qquad \mu \in \mathbb{R}, \sigma \in (0, \infty)$$

- This has a characteristic pdf $f_{\mathcal{X}}$:

$$f_{\mathcal{X}}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = \mu$$

$$\text{Var}_{\mathcal{D}}[\mathcal{X}] = \sigma^2$$

Gaussian Distribution

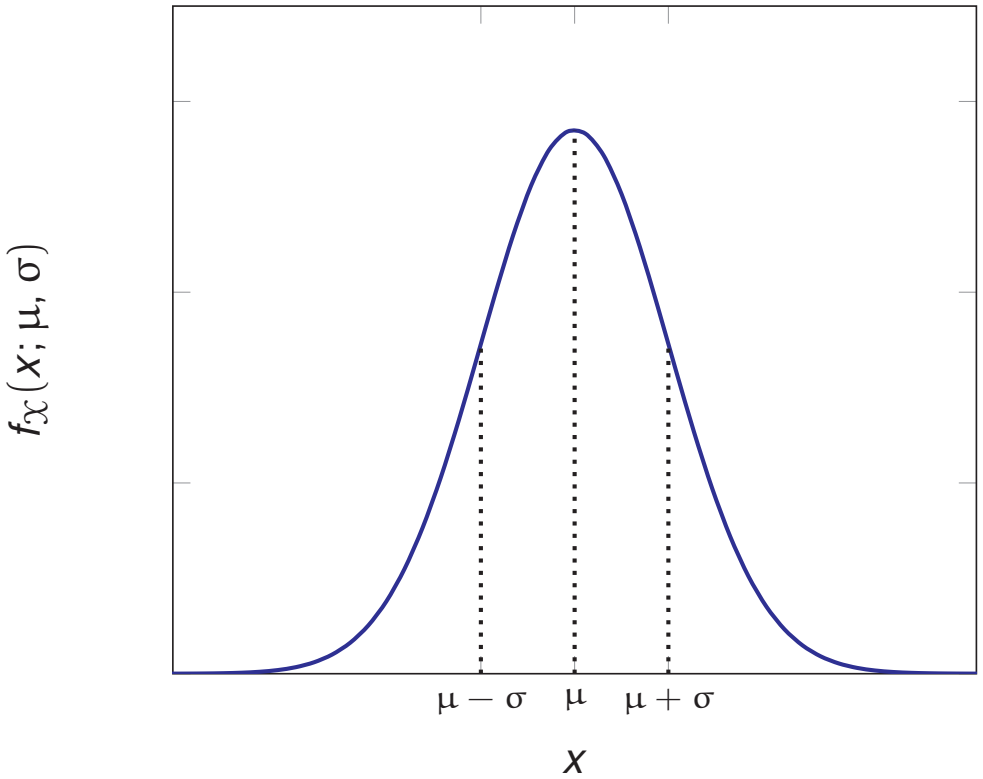
- And a cdf, F_X:

$$F_X(x;\mu,\sigma) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{t^2}{2}}\,dt$$
$$= \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

Where:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt$$

Gaussian Distribution



Gaussian Distribution: Properties

- **The Standard Normal Distribution**
  For any Gaussian distributed random variable, $\mathcal{X}$, with outcomes, $x \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to a standard normal distributed random variable, $\mathcal{Z}$, with outcomes, $z \sim \mathcal{N}(0, 1)$, by the following transformation:

$$\mathcal{Z} = \frac{\mathcal{X} - \mu}{\sigma}$$

Gaussian Distribution: Properties

- **Central Limit Theorem**
  Let $\{x_i\}_{i=1}^n$ be a random sample drawn from a sequence of i.i.d. random variables $\{\mathcal{X}_i\}_{i=1}^n$.
  Each $\mathcal{X}_i$ has the same probability distribution, with expectation μ and variance σ^2.
  Consider the random variable $\mathcal{Z} = \frac{\frac{1}{n}\sum_{i=1}^n \mathcal{X}_i - \mu}{\sigma/\sqrt{n}}$, with outcomes $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$:

$$\text{As:} \quad n \longrightarrow \infty$$
$$\text{Then:} \quad z \sim \mathcal{N}(0,1)$$

Multivariate Gaussian Distribution

- A random vector $\mathcal{X} = [\mathcal{X}_1, ..., \mathcal{X}_n]^T$, with outcomes $\mathbf{x}$, is distributed like a **multivariate Gaussian** distribution if it has the following characterisation:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \text{where:} \qquad \boldsymbol{\mu} \in \mathbb{R}^n,\ \boldsymbol{\Sigma}^T = \boldsymbol{\Sigma},\ \boldsymbol{\Sigma} \succ 0$$

- This has a characteristic pdf $f_{\mathcal{X}}$:

$$f_{\mathcal{X}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = \boldsymbol{\mu}$$

$$\text{Cov}_{\mathcal{D}}[\mathcal{X}] = \boldsymbol{\Sigma}$$

Multivariate Gaussian Distribution: Properties

- **Correlation & Independence**
  If any two elements within a multivariate Gaussian distribution are **uncorrelated** then they are **independent**

- **Marginal Distribution**
  To obtain the marginal distribution over a subset of random variables within a multivariate Gaussian we need only drop the irrelevant variables from the mean vector and the covariance matrix

- **Affine Transformation**
  For random variables $\mathcal{Y}$, $\mathcal{X}$, with outcomes $\mathbf{y}$, $\mathbf{x}$, respectively, then if $\mathcal{Y} = \mathbf{c} + \mathbf{B}\mathcal{X}$, where $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, then:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{c} + \mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$$

Multivariate Gaussian Distribution: Properties

- **Marginal & Conditional Distributions of Linear Gaussian Models**

  Given a Gaussian marginal distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

  Where: $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{L} \in \mathbb{R}^{m \times m}$

  Then:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T)$$

$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\Sigma}[\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}], \boldsymbol{\Sigma})$$

  Where: $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$

Multivariate Gaussian Distribution: Isocontours

- Note that the isocontours of a multivariate Gaussian are characterised by the exponent of the pdf:

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \text{const.}$$

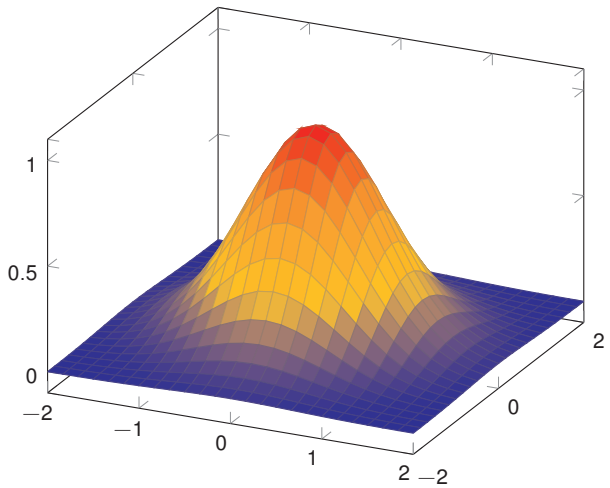
- But this is nothing more than the equation for the isocontours of the quadratic form of a positive definite matrix that we discussed in the last lecture

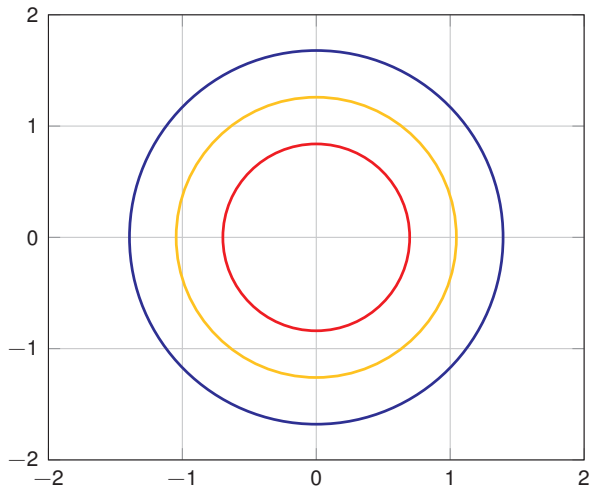
Multivariate Gaussian Distribution: Isocontours

- Using this result we can see that the **isocontours** are **ellipsoids** with:
  - **centre** at μ
  - **axes** pointing in the direction of the eigenvectors of Σ^{-1}
  - **radii** proportional to the inverse square root of the corresponding eigenvalues of Σ^{-1}
    or alternatively:
    proportional to the square root of the corresponding eigenvalues of Σ

Bivariate Gaussian: Isotropic

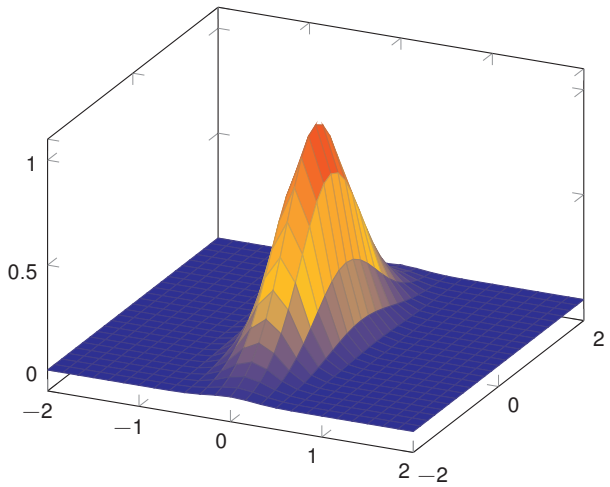
$$\Sigma = \sigma^2 \mathbf{I} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

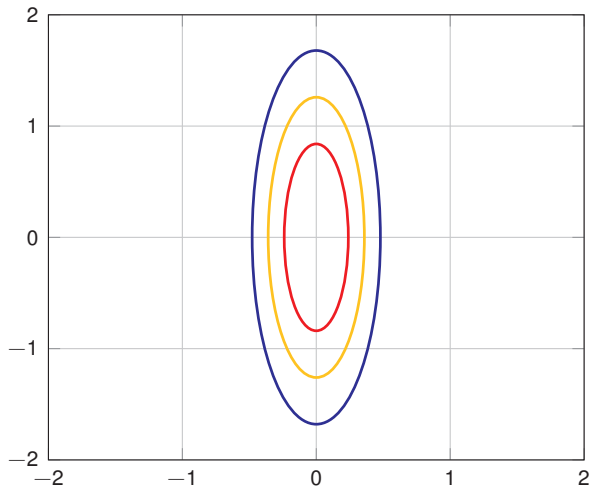




Bivariate Gaussian: Anisotropic

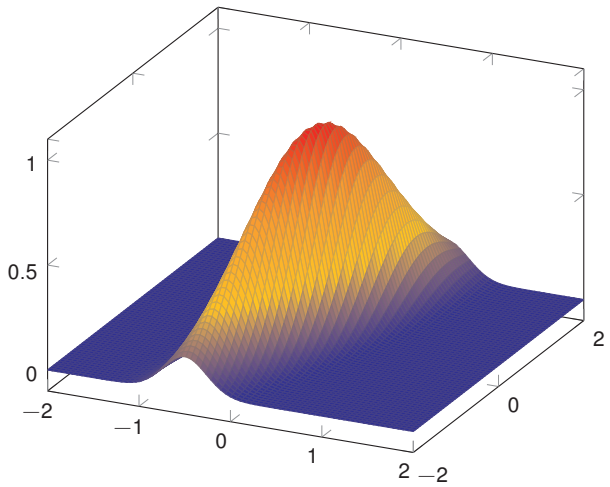
$$\Sigma = \text{diag}\left(\sigma_1^2, \sigma_2^2\right) = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

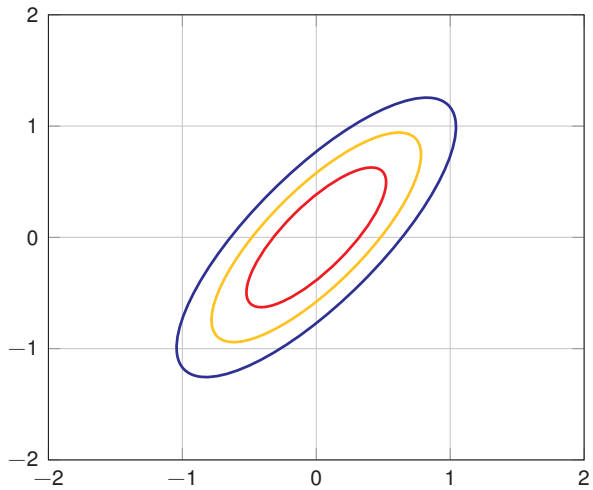




Bivariate Gaussian: Fully General

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_2, x_1) & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}$$





Beta Distribution

- Let $\mathcal{X}$ be a continuous random variable, taking values $x \in [0, 1]$, such that $\mathcal{D}$ follows a **Beta** distribution:

$$x \sim \text{Beta}(a, b) \qquad \text{where:} \qquad a, b > 0$$

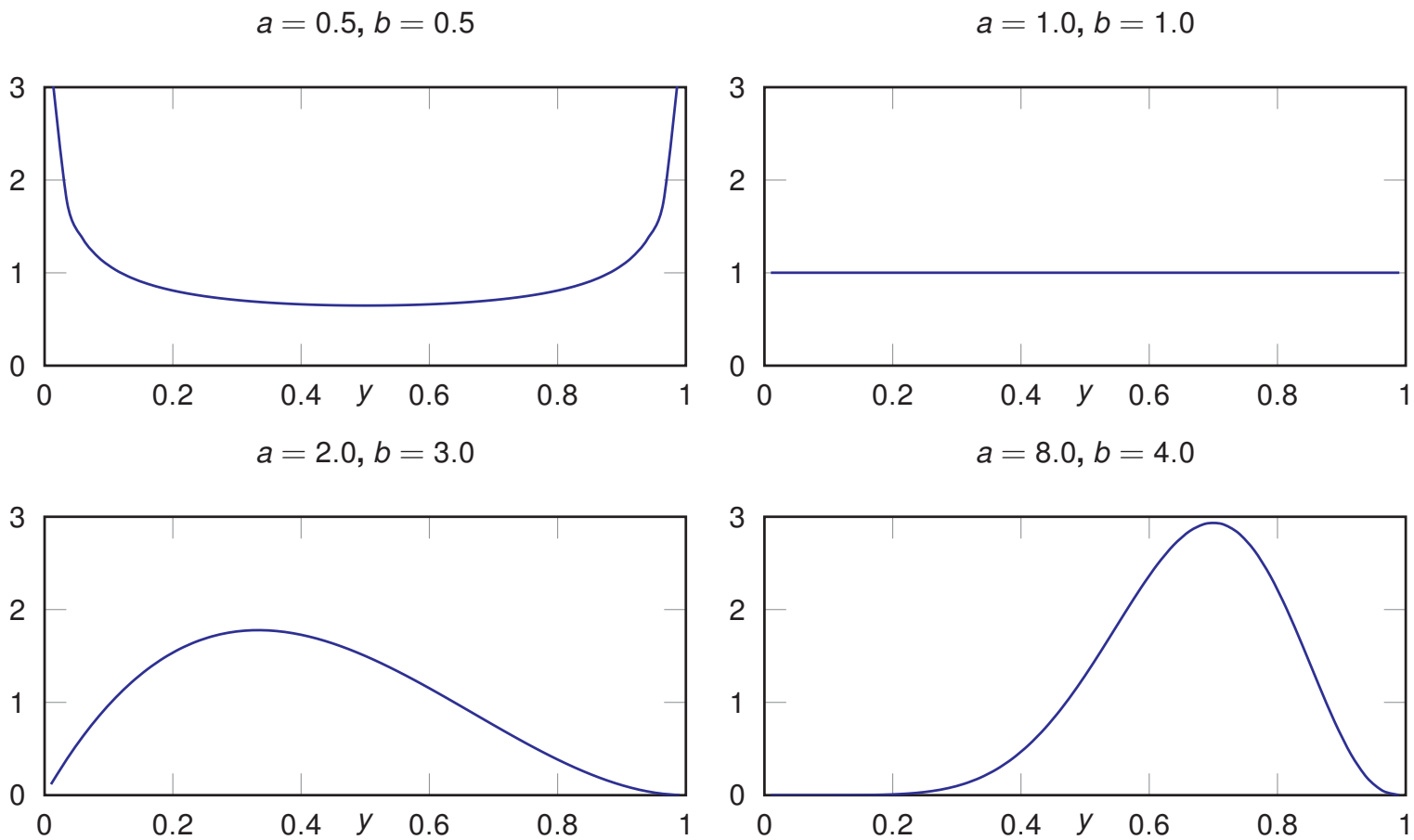
- This has a characteristic pdf $f_{\mathcal{X}}$:

$$f_{\mathcal{X}}(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad \text{where:} \quad \Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

$$\mathbb{E}_{\mathcal{D}}[\mathcal{X}] = \frac{a}{a+b}$$

$$\text{Var}_{\mathcal{D}}[\mathcal{X}] = \frac{ab}{(a+b)^2(a+b+1)}$$

Beta Distribution: Examples



Lecture Overview

Jensen's Inequality

- **Jensen's inequality** relates the value of a convex function of an integral (or sum) to the integral (or sum) of a convex function
- Since integrals often appear in our calculation of expectations then a probabilistic version of Jensen's inequality follows quite naturally

Jensen's Inequality: Probabilistic Version

- For a convex function, f, and a random variable, $\mathcal{X}$, with probability distribution $\mathcal{D}$:

$$\mathbb{E}_{\mathcal{D}}[f(\mathcal{X})] \geqslant f(\mathbb{E}_{\mathcal{D}}[\mathcal{X}])$$

- For example, $f(x) = -\log(x)$ is convex since:

$$\nabla^2_x(-\log(x)) = \frac{1}{x^2} \geqslant 0$$

So:

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}}[-\log(\mathcal{X})] \geqslant -\log(\mathbb{E}_{\mathcal{D}}[\mathcal{X}]) \\ \implies & \mathbb{E}_{\mathcal{D}}[\log(\mathcal{X})] \leqslant \log(\mathbb{E}_{\mathcal{D}}[\mathcal{X}]) \end{aligned}$$

Jensen's Inequality: Proof

- **Proof:**

We prove the finite form (for sums) only:

If f is a convex function and $\{\lambda_i \in \mathbb{R}^+\}_{i=1}^n$ are a set of non-negative real numbers such that $\sum_{i=1}^n \lambda_i = 1$, then $\forall x_i$:

$$
\begin{aligned}
f\left(\sum_{i=1}^{n} \lambda_i x_i\right) &= f\left(\lambda_1 x_1 + (1-\lambda_1)\sum_{i=2}^{n} \frac{\lambda_i}{1-\lambda_1} x_i\right) \\
&\leqslant \lambda_1 f(x_1) + (1-\lambda_1) f\left(\sum_{i=2}^{n} \frac{\lambda_i}{1-\lambda_1} x_i\right) \quad \textbf{By definition of convexity} \\
&\leqslant \sum_{i=1}^{n} \lambda_i f(x_i)
\end{aligned}
$$

Where the last line follows by induction, since: $\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} = 1$

If we let $\{\lambda_i = p_{\mathcal{X}}(x_i)\}_{i=1}^n$, where $p_{\mathcal{X}}(\cdot)$ is the pmf associated with some random variable $\mathcal{X}$, with outcomes $\{x_i\}_{i=1}^n$, then the proof follows.

Lecture Overview

Concentration Inequalities

- **Concentration inequalities** provide probabilistic bounds on how a random variable deviates from some value
- They are often used in theoretical machine learning to prove the efficacy of certain algortihms

Markov's Inequality

- Let $\mathcal{X}$ be a random variable which is non-negative, with expectation μ.
  Then, for every constant $a > 0$:

$$\mathbb{P}(\mathcal{X} \geqslant a) \leqslant \frac{\mu}{a}$$

Chebyshev's Inequality

- Let $\mathcal{X}$ be a random variable, for which the expectation, μ, and variance, σ^2, are finite.
  Then, for every constant $a > 0$:

$$\mathbb{P}(|\mathcal{X} - \mu| \geqslant a) \leqslant \frac{\sigma^2}{a^2}$$

Hoeffding's Inequality

- Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be independent random variables, such that each $\mathcal{X}_i$ is bounded by the interval $[a_i, b_i]$.
  Let the empirical mean of these random variables be defined as: $\overline{\mathcal{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i$.
  Then:

$$\mathbb{P}\left(\overline{\mathcal{X}} - \mathbb{E}[\overline{\mathcal{X}}] \geqslant t\right) \leqslant \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

  And:

$$\mathbb{P}\left(|\overline{\mathcal{X}} - \mathbb{E}[\overline{\mathcal{X}}]| \geqslant t\right) \leqslant 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

  Where $t \geqslant 0$

Lecture Overview

Lecture Summary

- **Probability Theory** provides a framework for reasoning with uncertainty
- We have introduced the basic machinery for characterising uncertainty in a probabilistic setting
- We have also introduced some distributions and results that will be of direct use in machine learning