# Mathematics for Machine Learning

**Prof Willie Brink**

Applied Mathematics, Stellenbosch University

## Lecture 5: Probability and Distributions

# Contents of the module

## 6.1 Construction of a probability space

Sample space $\Omega$: set of all possible outcomes of an experiment.

Event: subset $A$ of the sample space.

Probability: a number $P(A)$ that measures the probability or degree of belief that event $A$ will occur when the experiment is executed, such that:

- $0 \leq P(A) \leq 1$ for any event $A$
- $P(\Omega) = 1$
- $P(A \cup B) = P(A) + P(B)$ for mutually exclusive events $A$ and $B$

A random variable $X$ is defined by a set of possible values (or states), and probabilities associated with elements or subsets of that set.

# 6.2 Discrete and continuous probabilities

## Discrete random variables

A discrete random variable $X$ has probability mass function $p(x) = P(X = x)$.

Bivariate mass function visualised as a probability table:

- joint probability: $p(x, y) = P(X = x \text{ and } Y = y)$

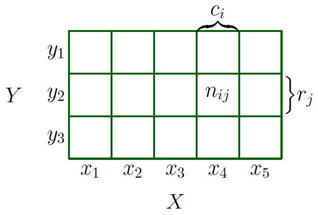
  an element in the probability table

- marginal probability: $p(x)$

  column sum, or row sum for $p(y)$

- conditional probability: $p(x \mid y)$

  fraction of an element within its row, or within its column for $p(y|x)$



## Continuous random variables

A continuous random variable $X$ is defined on the real line $\mathbb{R}$, with a probability density function $f(x)$, such that $f(x) \geq 0$ for all $x \in \mathbb{R}$, and

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx \;, \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1 \;.$$

The cumulative distribution function of $X$: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(z)\,dz$

Note that $P(X = a) = \int_a^a f(x)\,dx = 0$.

The probability that a continuous random variable will assume any fixed value is zero.

But $P(a - \frac{1}{2}\epsilon \leq X \leq a + \frac{1}{2}\epsilon) \approx \epsilon\, f(a)$, so $f(a)$ gives an indication of how relatively likely it is that $X$ is near $a$.

## 6.3 Sum rule, product rule, and Bayes' theorem

Sum rule: $p(x) = \begin{cases} \sum_y p(x,y) & \text{if } y \text{ is discrete} \\ \int_y p(x,y)\,dy & \text{if } y \text{ is continuous} \end{cases}$

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

Bayes' theorem: $p(x|y) = \dfrac{p(y|x)p(x)}{p(y)}$

$p(x)$ is called the prior, $p(y|x)$ is the likelihood, $p(x|y)$ is the posterior, and $p(y)$ is the evidence usually computed as $\sum_x p(y|x)p(x)$ or $\int p(y|x)p(x)\,dx$

A random variable can be multivariate; then we would write $\boldsymbol{x}$ and $\boldsymbol{y}$ in the above.

## Exercise 6.4, p. 222

Bag A has 4 mangos and 2 apples. Bag B has 4 mangos and 4 apples.

If a biased coin (probability of heads 0.6) lands on heads, we pick a fruit from bag A. If it lands on tails, we pick a fruit from bag B.

Your friend flips the coin (you can't see), and picks a fruit. It is a mango. What is the probability that it comes from bag B?

Let $h$ be the event that the coin lands on heads (bag A), and $t$ for tails (bag B).

Let $m$ be the event that the chosen fruit is a mango.

We are given: $p(h) = 0.6, \;\; p(t) = 0.4, \;\; p(m|h) = \frac{4}{6}, \;\; p(m|t) = \frac{4}{8}$

Then $p(t|m) = \frac{p(m|t)p(t)}{p(m)} = \frac{p(m|t)p(t)}{p(m,t) + p(m,h)} = \frac{p(m|t)p(t)}{p(m|t)p(t) + p(m|h)p(h)} = \frac{1}{3}\,.$

## 6.4 Summary statistics and independence

The expected value of a function $g$ of a random variable $X \sim p(x)$:

$$\mathbb{E}[g(x)] = \sum_x g(x)p(x) \quad \text{or} \quad \mathbb{E}[g(x)] = \int g(x)p(x)\,dx$$

The mean (or average) of $X$ is $\mathbb{E}[x]$.

A one-dimensional random variable also has a median and one or more modes.

Note the expected value is linear: $\mathbb{E}[ag(x) + bh(x)] = a\mathbb{E}[g(x)] + b\mathbb{E}[h(x)]$

The variance of $X$ with mean $\mu$: $\mathbb{V}[x] = \mathbb{E}[(x-\mu)^2] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$

The covariance between two univariate random variables $X$ and $Y$:

$$\mathrm{Cov}[x, y] = \mathbb{E}_{X,Y}\big[\,(x - \mathbb{E}_X[x])\,(y - \mathbb{E}_Y[y])\,\big]$$

The covariance matrix of a multivariate random variable $X$ with mean $\boldsymbol{\mu}$:

$$\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}] = \mathbb{V}[\boldsymbol{x}] = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^{\mathsf{T}}] = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^T$$

$$= \begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \cdots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \mathrm{Cov}[x_D, x_2] & \cdots & \mathrm{Cov}[x_D, x_D] \end{bmatrix}$$

The covariance matrix is symmetric and positive semidefinite (usually positive definite), and gives an indication of the spread of the data.

The correlation between $X$ and $Y$: $\mathrm{corr}[x, y] = \dfrac{\mathrm{Cov}[x, y]}{\mathbb{V}[x]\mathbb{V}[y]} \in [-1, 1]$

Positive correlation $\mathrm{corr}[x, y]$ means that when $x$ grows, $y$ is expected to grow.
Negative correlation means that as $x$ increases, $y$ decreases.

The empirical mean and covariance of $N$ observations of $X$:

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \quad \text{and} \quad \boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^{\mathsf{T}}$$

## Statistical independence

Two random variables $X$ and $Y$ are independent if and only if $p(x, y) = p(x)p(y)$.

If $X$ and $Y$ are independent, then $p(x \mid y) = p(x)$.

Two random variables $X$ and $Y$ are conditionally independent given $Z$, if and only if
$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)$.

If $X$ and $Y$ are conditionally independent given $Z$, then $p(x \mid y, z) = p(x \mid z)$.

# 6.5 Gaussian distribution

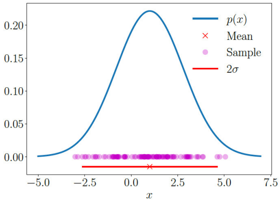
The univariate Gaussian (or normal) distribution: $p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

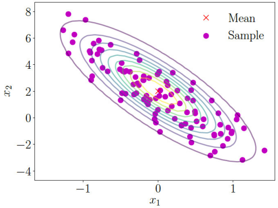
The multivariate Gaussian (or normal) distribution, with $\boldsymbol{x} \in \mathbb{R}^D$:

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

We often write $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Widely used as it has closed-form expressions for marginal and conditional distributions.



Let's write the Gaussian distribution in terms of a concatenation of states $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right)$$

Marginal: $p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y} = \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$

Conditional: $p(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})$ where $\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$$

The product of two Gaussians $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{a}, \boldsymbol{A})$ and $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{b}, \boldsymbol{B})$ is $c\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{c}, \boldsymbol{C})$, where

$$\boldsymbol{C} = \left(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1}\right)^{-1}, \quad \boldsymbol{c} = \boldsymbol{C}\left(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}\right),$$

$$c = (2\pi)^{-\frac{D}{2}}|\boldsymbol{A} + \boldsymbol{B}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^{\mathsf{T}}(\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\right)$$

## Sums and linear transformations

If $X$ and $Y$ are independent Gaussian variables, then $p(\boldsymbol{x}+\boldsymbol{y}) = \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y)$.

If $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $Y$ has states $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, then $p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}})$.

## Sampling from multivariate Gaussian distributions

Suppose we want to generate samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1. source uniformly random samples in [0,1];
2. apply Box-Müller transform to obtain samples from a univariate Gaussian;
3. collate a vector of these samples to obtain a sample from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Now, if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\mu}$ where $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{\Sigma}$, is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Can use Cholesky decomposition to find $\boldsymbol{A}$ from $\boldsymbol{\Sigma}$.

## 6.6 Conjugacy and the exponential family

Bernoulli distribution with parameter $\mu \in [0,1]$: $p(x \,|\, \mu) = \mu^x (1-\mu)^{1-x}$, $x \in \{0,1\}$

The binomial distribution describes the probability of observing $m$ occurences of $X = 1$ in $N$ samples from a Bernoulli distribution where $p(X = 1) = \mu$. Hence

$$p(m \,|\, N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

The beta distribution over a continuous random variable $\mu \in [0,1]$ (e.g. the parameter of a Bernoulli distribution) has two parameters $\alpha > 0$ and $\beta > 0$:

$$p(\mu \,|\, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{1-\alpha} (1-\mu)^{1-\beta}$$ where $\Gamma(\cdot)$ is the Gamma function

Intuitively, $\alpha$ moves the probability mass towards 0, and $\beta$ moves it towards 1.

## Conjugacy

According to Bayes, the posterior is proportional to the prior times the likelihood.

A prior is conjugate for the likelihood if the posterior is of the same type as the prior.
Convenient, as we can calculate the posterior by updating the parameters of the prior.

**Example:** Let $X$ be a binomial random variable with parameters $N$ and $\mu$ (number of heads in $N$ flips of a biased coin with $\mu$ the probability of heads in one flip). We place a beta prior on $\mu$, with parameters $\alpha$ and $\beta$, and then observe some outcome $x = h$ (we see $h$ heads in $N$ flips).

Posterior on $\mu$: $p(\mu \mid x = h) \propto p(x \mid N, \mu)\, p(\mu \mid \alpha, \beta) \propto \mu^{h+\alpha-1}(1-\mu)^{(N-h)+\beta-1}$

...which is a beta distribution with parameters $h + \alpha$ and $N - h + \beta$.

The beta prior is conjugate for the parameter $\mu$ in the binomial likelihood function.

## Exponential family

The Gaussian distribution is a member of the exponential family.

This family of distributions, parameterised by $\boldsymbol{\theta} \in \mathbb{R}^D$, has the form

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\big(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\big)$$

$h(\boldsymbol{x})$ can be absorbed into the exponent by adding $\log h(\boldsymbol{x})$ to $\phi(\boldsymbol{x})$, and $\exp\big(-A(\boldsymbol{\theta})\big)$ is the normalisation constant, so that

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \exp\big(\boldsymbol{\theta}^T \phi(\boldsymbol{x})\big)$$

For the univariate Gaussian $\mathcal{N}(\mu, \sigma^2)$, we'd have $\phi(x) = [x, x^2]^T$ and $\boldsymbol{\theta} = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]^T$.

The Bernoulli distribution with parameter $\mu$ is also a member of this family, where $h(x) = 1,\ \phi(x) = x,\ \boldsymbol{\theta} = \log \frac{\mu}{1-\mu},$ and $A(\theta) = \log(1 + \exp(\theta))$.

# 6.7 Change of variables

Let $X$ be a continuous random variable with pdf $f(x)$. What is the pdf of $Y = U(X)$?

## Distribution function technique

$$F_Y(y) = P(Y \leq y) = P(U(X) \leq y) = P(X \leq U^{-1}(y)) = F_X(U^{-1}(y))$$

To obtain the pdf of $Y$, we differentiate its cdf: $f(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(U^{-1}(y))$

### Probability integral transform

If $X$ has a strictly monotonic cdf $F_X$, then $Y = F_X(X)$ has a uniform distribution.

We can use this result to sample from the distribution of $X$, by transforming uniformly random samples in $[0, 1]$ with the inverse cdf $(F_X)^{-1}$.