Week 2: Estimators and Common Distributions

Probability Theory for Machine Learning

Dr Daniel Worrall

AMLab, University of Amsterdam

September 4, 2019

Recap

Last week, we examined sample spaces Ω, event spaces $\mathcal{F}$, and *probability measures*[1] $P : \mathcal{F} \rightarrow [0, 1]$.

> *The triple* $(\Omega, \mathcal{F}, P)$ *defines a probability space.*

Given a probability space, we are now in position to measure certain properties. These properties will hopefully correspond to real-life, ‘measurable’ quantities.

[1] Probability measure is the catch-all term for PMF and PDF.

I: Random Variables

Random variables

In science, we do not work with events, but rather, we are given a (finite) number of *samples* or *observations* $\mathcal{D} = \{x_1, x_2, ..., x_N\}$, whose distribution is $P(x)$.

It is the job of statisticians to recover $P(x)$ from the dataset $\mathcal{D}$.

$$\{x_1, x_2, ..., x_N\} \xrightarrow{\text{something clever}} P(x)$$

Before we do that, we are going to study how to *sample* from $P(x)$

$$P(x) \xrightarrow{\text{sampling}} \{x_1, x_2, ..., x_N\}$$

We will then study the behaviour of functions of these samples, so that we can construct functions, which *measure* useful properties of $P(x)$.

Random variables

Random variable

A random variable $X : \Omega \to \mathcal{X}$ is a map from sample space Ω to numbers[2] $\mathcal{X}$.

e.g. The sample space for a single coin is $\Omega = \{\texttt{head}, \texttt{tail}\}$. We can map this to the real numbers as such

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \texttt{heads} \\ 0, & \text{if } \omega = \texttt{tails} \end{cases}$$

e.g. The number of hot dogs n I eat in a hot-dog eating competition has sample space $\Omega = \{0, 1, 2, 3, ...\}$. The RV X for the number of hot dogs I *do* eat is

$$X(\omega_n) = n.$$

Random variables are useful, because we can do maths with them e.g. $0 + 1$, when the corresponding operations make no sense on sample space e.g. $\texttt{heads} + \texttt{tails}$

[2] Nit-picky statisticians say that a random variable is a map from sample space to a 'measurable space', a technicality we will not go into.

Random variables

Random variables can also be more complicated functions of the sample spaces.

e.g. Consider throwing two dice.

$$\Omega = \{(1,1),(1,2),\dots,(2,1),(2,2),\dots,(6,6)\}$$

We can define multiple random variables:

- X The sum of eyes on both dice.
- Y The product of the number of eyes on each die.
- Z The number of eyes on the first die

For $\omega = (2,3)$, the random variables take the following values:

- $X(\omega) = 5$, $Y(\omega) = 6$, $Z(\omega) = 2$

Random variables

The probability that random variable X takes on value x is

$$P(X = x)$$

A random variable only takes on a value *after* an experiment has taken place.

e.g. In our hot dog eating example I may have $P(X = 3) = 0.5$.

The notation

$$X \sim P(x)$$

is used to denote that random variable X is *distributed* according to $p(x)$.

Machine Learning Notation
The notation $P(X = x)$ is explicit, but machine learning people are a bit sloppy, writing $P(x)$ instead. Others prefer to write $P_X(x)$, it really depends on your preference.

To make things more confusing, machine learning people sometimes also refer to x as the random variable, instead of X!

We are going to be true machine learners and adopt this (lazy) convention. Furthermore, we will write $P(x)$ instead of $P(X = x)$.

II: Maths With Random Variables

Sample Means

Say I roll an unbiased die. We represent the values of its sides with random variable x. When I roll it 5 times, the resulting samples are $\{4, 2, 6, 2, 1\}$. The average of the samples, called the *sample mean* $\bar{x}$ is

$$\bar{x} = \frac{4+2+6+2+1}{5} = 3.$$

In general, for samples $\{x_1, x_2, ..., x_N\}$, the sample mean $\bar{x}$ is

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_i$$

Note that the *sample* mean is a function of the *samples*! For different sets of samples, the sample mean is different.

If someone gave you the PMF $P(x)$, what sample mean would you *expect*?

Expected mean

Say I now roll the die N times. Denoting the number of times we see the number 1 as N_1, 2 as N_2, etc., the general formula for die averages is

$$\bar{x}(N) = \frac{1 \cdot N_1 + 2 \cdot N_2 + 3 \cdot N_3 + 4 \cdot N_4 + 5 \cdot N_5 + 6 \cdot N_6}{N}$$

Let's take the limit $N \to \infty$,

$$\begin{aligned} \bar{x}(\infty) &= \lim_{N \to \infty} \left[1 \cdot \frac{N_1}{N} + 2 \cdot \frac{N_2}{N} + 3 \cdot \frac{N_3}{N} + 4 \cdot \frac{N_4}{N} + 5 \cdot \frac{N_5}{N} + 6 \cdot \frac{N_6}{N} \right] \\ &= 1 \cdot P_1 + 2 \cdot P_2 + 3 \cdot P_3 + 4 \cdot P_4 + 5 \cdot P_5 + 6 \cdot P_6 \end{aligned}$$

We have that $\lim_{N \to \infty} \frac{N_i}{N} = P_i$ from the frequentist definition of probability.

Expected Means

Expected means
Introducing the symbol $\mu := \bar{x}(\infty)$, we have that

$$\boxed{\mu = \sum_{x \in \mathcal{X}} x P(x) \quad \text{or} \quad \mu = \int_{\mathcal{X}} x p(x) \, \mathrm{d}x.}$$

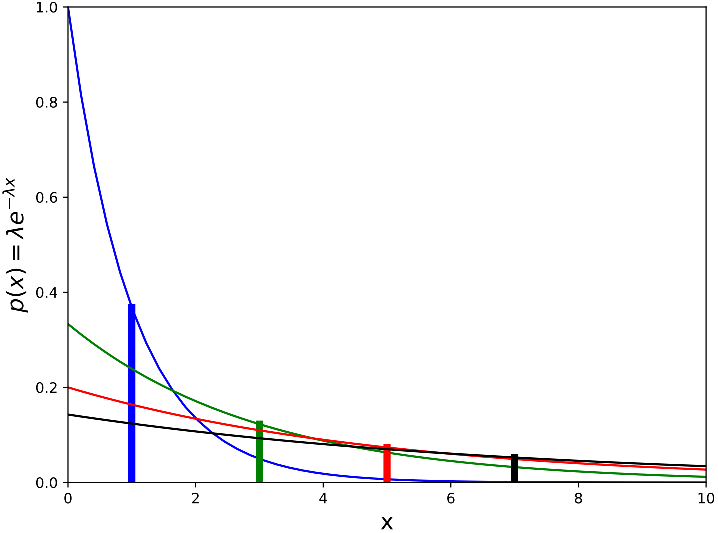
e.g. Waiting times t are modelled by an exponential distribution

$$p(t; \lambda) = \lambda e^{-\lambda t}.$$

The mean waiting time of $p(t; \lambda)$ is [HINT: $\int_0^\infty x e^{-x} \, \mathrm{d}x = 1$]

$$\mu = \int_0^\infty x \cdot \lambda e^{-\lambda x} \, \mathrm{d}x = \frac{1}{\lambda} \int_0^\infty y e^{-y} \, \mathrm{d}y = \frac{1}{\lambda}.$$

Means



Means

e.g. A computer vision system is used to sort out carrots on a conveyor belt according to their size x cm. Long carrots ($x > 10$ cm) are sold directly to supermarkets, all other carrots are used to make soup, which is then sold.
The distribution of carrot lengths is known to follow a uniform distribution

$$p(x) = U(x; 6, 20) = \frac{1}{14} \mathbb{I}\left[x \in [6, 20]\right]$$

What is the mean length of carrots not turned into soup?

Answer We know the distribution of carrots larger than 10 cm is uniform, with

$$p(x|x > 10) = U(x; 10, 20)$$

so the mean is

$$\int_{10}^{20} x \cdot \frac{1}{20 - 10} \,\mathrm{d}x = \frac{1}{10} \left[\frac{x^2}{2}\right]_{10}^{20} = \frac{1}{10} \left(\frac{20^2}{2} - \frac{10^2}{2}\right) = 15$$

Expectations

Expectations
Say instead of taking the mean of x, we took the mean of x^2? What would be the formula for the expected squared mean? Quite simply

$$\overline{x^2}(N) = \frac{1^2 \cdot N_1 + 2^2 \cdot N_2 + 3^2 \cdot N_3 + 4^2 \cdot N_4 + 5^2 \cdot N_5 + 6^2 \cdot N_6}{N}$$

Taking limit $N \to \infty$,

$$\overline{x^2}(\infty) = 1^2 \cdot P_1 + 2^2 \cdot P_2 + 3^2 \cdot P_3 + 4^2 \cdot P_4 + 5^2 \cdot P_5 + 6^2 \cdot P_6$$

so in general

$$\mu = \sum_{x \in \mathcal{X}} x^2 P(x) \quad \text{or} \quad \mu = \int_{\mathcal{X}} x^2 p(x) \, \mathrm{d}x.$$

Expectations

Expectations

The *expectation* of a function $f : \mathcal{X} \to \mathcal{Y}$ of a random variable x is

$$\boxed{\mathbb{E}_x[f(x)] = \sum_x f(x)P(x) \quad \text{or} \quad \mathbb{E}_x[f(x)] = \int f(x)p(x)\,\mathrm{d}x.}$$

Sometimes we also see the notation $\mathbb{E}_x[f]$, $\mathbb{E}_{x\sim P(x)}[f]$, $\mathbb{E}_P[f]$, or $\mathbb{E}_{P(x)}[f]$. We just use $\mathbb{E}[f(x)]$ instead of $\mathbb{E}_x[f(x)]$, when it is obvious.

Think of passing samples of $p(x)$ through f and taking the mean in $\mathcal{Y}$.

The mean function is just the expectation of the random variable x,

$$\mu = \mathbb{E}[x]$$

Expectations

e.g. What is $\mathbb{E}[x^3]$ where $p(x) = U(x; 0, 1)$

$$\begin{aligned}\mathbb{E}[x^3] &= \int_0^1 x^3 p(x)\,\mathrm{d}x \\ &= \int_0^1 x^3 \cdot 1\,\mathrm{d}x \\ &= \left[\frac{1}{4}x^4\right]_0^1 \\ &= \frac{1}{4}\end{aligned}$$

Expectations

Expectations are *linear operators*, so

$$\mathbb{E}[ax+b] = a\mathbb{E}[x] + b.$$

This follows directly from the linearity of the summation/integral

$$\begin{aligned}
\mathbb{E}[ax+b] &= \sum_{x\in\mathcal{X}} (ax+b)\cdot P(x) \\
&= \sum_{x\in\mathcal{X}} a\cdot xP(x) + b\cdot P(x) \\
&= \sum_{x\in\mathcal{X}} a\cdot xP(x) + \sum_{x\in\mathcal{X}} bP(x) \\
&= a\underbrace{\sum_{x\in\mathcal{X}} xP(x)}_{=\mathbb{E}[x]} + b\underbrace{\sum_{x\in\mathcal{X}} P(x)}_{=1} \\
&= a\mathbb{E}[x] + b
\end{aligned}$$

Variance

Variance
The variance $\mathbb{V}$ of x is defined as

$$\boxed{\mathbb{V}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \underbrace{\mathbb{E}[x^2] - \mathbb{E}[x]^2}_{\text{in the werkcollege}}.}$$

It measures the average squared distance of x from $\mathbb{E}[x]$, or squared spread.

e.g. The variance of the exponential distribution is [Hint: integrate by parts]

$$\mathbb{E}[t^2] = \int_0^\infty t^2 \lambda e^{-\lambda t}\, \mathrm{d}t = \underbrace{\left[-t^2 e^{-\lambda t}\right]_0^\infty}_{=0} + \int_0^\infty 2te^{-\lambda t}\, \mathrm{d}t$$

$$= \frac{2}{\lambda}\int_0^\infty t\lambda e^{-\lambda t}\, \mathrm{d}t = \frac{2}{\lambda}\mathbb{E}[t] = \frac{2}{\lambda}\frac{1}{\lambda} = \frac{2}{\lambda^2}$$

$$\mathbb{V}[t] = \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Variance

Variance

A useful identity to know is

$$\mathbb{V}[ax+b] = a^2\mathbb{V}[x]$$

because

$$\begin{aligned}
\mathbb{V}[ax+b] &= \mathbb{E}[((ax+b) - \mathbb{E}[ax+b])^2] \\
&= \mathbb{E}[(ax+b-a\mathbb{E}[x]-b)^2] \\
&= \mathbb{E}[(ax-a\mathbb{E}[x])^2] \\
&= \mathbb{E}[a^2(x-\mathbb{E}[x])^2] \\
&= a^2\mathbb{E}[(x-\mathbb{E}[x])^2] \\
&= a^2\mathbb{V}[x]
\end{aligned}$$

Notice how the variance is invariant under shifts of x. Notice also that we get a^2, because the variance measures the average *squared* spread of $p(x)$.

Standard deviation

Variance measures the squared *spread* or *width* of a distribution.

Standard deviation

The standard deviation σ is defined as the square root of the variance. It is a measure of how spread out probability distribution p it about its mean.

$$\boxed{\sigma^2 = \mathbb{V}[x].}$$

A nice consequence of using the standard deviation is that it scales linearly with a

$$\sqrt{\mathbb{V}[ax+b]} = \sqrt{a^2\mathbb{V}[x]} = a\sigma$$

e.g. The standard deviation of the exponential distribution is $\frac{1}{\lambda}$, which coincidentally is also its mean.

Moments

Moments

The n^{th} moment of a distribution about c is defined as

$$\mu_n = \mathbb{E}[(x-c)^n]$$

Examples: mean typically just written μ ($\mu = \mu_1 = \mathbb{E}[x]$), power ($\mu_2 = \mathbb{E}[x^2]$). If c is not given, then we just assume $c = 0$.

Central Moments

If we set $c = \mu$, then we have a *central moment* σ_n.

$$\sigma_n = \mathbb{E}[(x-\mu)^n]$$

Example: variance ($\sigma^2 = \sigma_2 = \mathbb{E}[(x-\mu)^2]$)

Normalised Moments

The n^{th}-normalised moment of a distribution is defined as

$$\frac{\mu_n}{\sigma^n} = \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^n\right]$$

where σ is the standard deviation. Examples: skewness ($n = 3$), kurtosis ($n = 4$), hyperskewness ($n = 5$), hyperflatness ($n = 6$)

n$^{\text{th}}$-moments of the Exponential Distribution

e.g. Find the n^{th} moment of the exponential $p(x;\lambda) = \lambda e^{-\lambda x}\mathbb{I}[x \geq 0]$.

Integrating by parts

$$\mu_n = \int_0^\infty t^n \lambda e^{-\lambda t}\,\mathrm{d}t = \underbrace{\left[-t^n e^{-\lambda t}\right]_0^\infty}_{=0} + \int_0^\infty n t^{n-1} e^{-\lambda t}\,\mathrm{d}t$$

$$= \frac{n}{\lambda}\int_0^\infty t^{n-1}\lambda e^{-\lambda t}\,\mathrm{d}t = \frac{n}{\lambda}\mu_{n-1}$$

Since we have the recurrence relation $\mu_n = \frac{n}{\lambda}\mu_{n-1}$ and we know that $\mu_1 = \frac{1}{\lambda}$

$$\boxed{\mu_n = \frac{n}{\lambda}\frac{n-1}{\lambda}\cdots\frac{2}{\lambda}\frac{1}{\lambda} = \frac{n!}{\lambda^n}}$$

Sample expectations

We derived the *mean value operator* from the *sample mean*. We then defined expectations, variances, and the standard deviation *for distributions*. Given a set of samples $\mathcal{D} = \{x_1, x_2, .., x_N\}$, can we reasonably guess these quantities?

We might guess something that looks right like:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2.$$

Quantities such as $\hat{\mu}$ and $\hat{\sigma}^2$ are called *estimators*. In the next section, we consider what makes a good estimator.

III: Common distributions

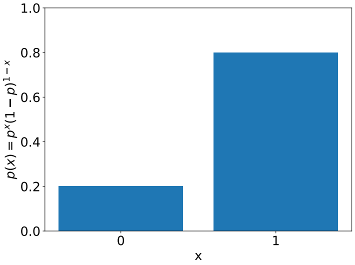
Bernoulli distributions

The simplest distribution is the *Bernoulli distribution*

$$x \sim \text{Ber}(x; p), \qquad \text{where } 0 \leq p \leq 1.$$

X is a binary random variable, with $P(X = 1) = p$, $P(X = 0) = 1 - p$. This is sometimes written more compactly

$$P(x) = p^x (1 - p)^{1-x}$$



Note p has two different meanings!! It is both a probability and parameter.

e.g. In a model of telecommunications model, x may be the random variable indicating whether a message is corrupted or not. $x \sim \text{Ber}(x; p)$ would be used to define the probability p that a message gets corrupted.

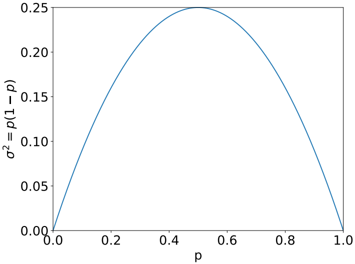
Bernoulli distribution

The mean of the Bernoulli is

$$\mu = \mathbb{E}[x] = \sum_x x P(x) = 0 \cdot (1-p) + 1 \cdot p = p$$

The variance of the Bernoulli is

$$\begin{aligned}\sigma^2 &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\ &= \sum_x x^2 P(x) - p^2 \\ &= 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 \\ &= p - p^2 = p(1-p)\end{aligned}$$



The variance is maximal when $p = 1/2$, with $\sigma^2 = 1/4$. What is the standard deviation?

Uniform distribution

The simplest continuous distribution is the *Uniform distribution*

$$x \sim \text{Uniform}(x; a, b) \qquad a < b$$

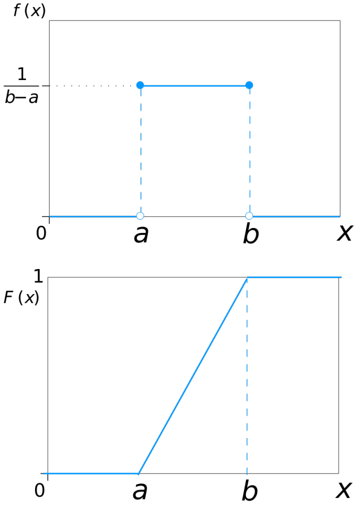
This density is flat on the interval $[a, b]$

$$\text{Uniform}(x; a, b) = \underbrace{\frac{1}{b-a}}_{\text{normalizer}} \mathbb{I}[x \in [a, b]]$$

You can work out the mean and variance

$$\mathbb{E}[x] = \frac{1}{2}(a + b)$$
$$\mathbb{V}[x] = \frac{1}{12}(b - a)^2$$

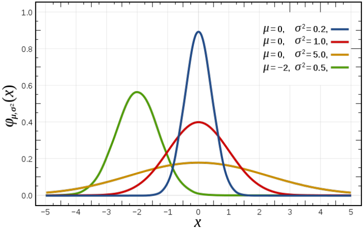
The CDF of the uniform density is a ramp.



Gaussian distribution

The distribution you are most likely to encounter is the *Normal* or *Gaussian distribution*

$$x \sim \mathcal{N}(x; \mu, \sigma^2) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{normalizer}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$



Commit this to memory!

It has 2 parameters, μ and σ^2. μ controls the location of the mode along the x axis and σ^2 controls the scale of the distribution.

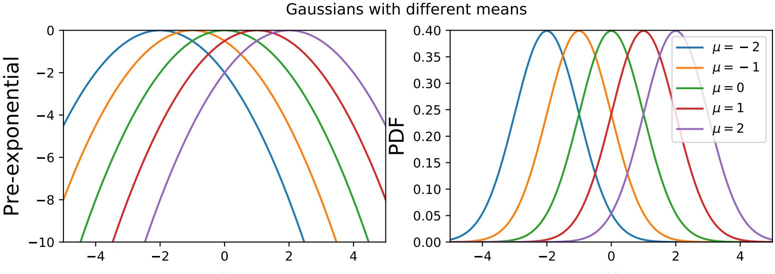
The Gaussian ccurs very often in nature due to the Central Limit Theorem (next week) and has some nice properties, which make it very easy to work with.

Gaussian distribution

Let's dissect the Gaussian. In essence, it is just an exponentiated quadratic.

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

The *location parameter* μ controls the centre of the Gaussian.

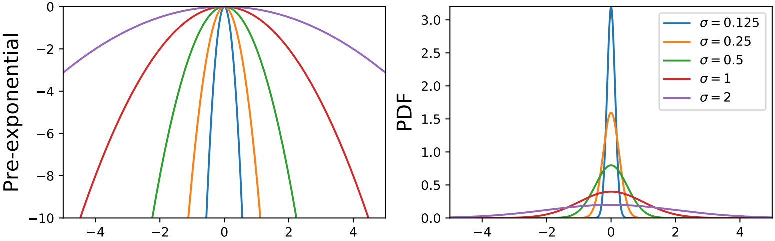


Gaussian distribution

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

The *scale parameter* σ controls the “width” of the Gaussian.

Gaussians with different variances



The exponential is needed to map the quadratics, which contain negative values, to positive numbers. The $\frac{1}{\sqrt{2\pi\sigma^2}}$ term then normalises the result.

Standard Gaussian distribution

A special case of the Gaussian is the *standard Gaussian*. It has a location parameter $\mu = 0$ and squared scale $\sigma^2 = 1$, so

$$\boxed{\mathcal{N}(x; 0, 1^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}}.$$

A useful identity is the derivative of this

$$\frac{\partial}{\partial x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} = -x \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}.$$

This makes the mean easy to compute

$$\int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} = \left[-\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}\right]_{-\infty}^{\infty} = 0.$$

Standard Gaussian distribution

We solve for the variance, by integrating by parts

$$\mathbb{V}[x] = \int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \mathrm{d}x$$

choosing

$$\begin{aligned} f(x) &= x && \implies f'(x) = 1 \\ g'(x) &= x \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} && \implies g(x) = -\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \end{aligned}$$

we get

$$\mathbb{V}[x] = \underbrace{\left[-x\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}\right]_{-\infty}^{\infty}}_{=0} + \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \mathrm{d}x}_{=\int p(x)\,\mathrm{d}x} = 1$$

Gaussian distribution

Using the substitution $y = \frac{x-\mu}{\sigma}$, we can find the mean and variance of $\mathcal{N}(x; \mu, \sigma^2)$ (note $x = \sigma y + \mu$ and $\mathrm{d}x/\sigma = \mathrm{d}y$):

$$\begin{aligned}
\mathbb{E}[x] &= \int x \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\
&= \mu + \sigma \underbrace{\int y \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \mathrm{d}y}_{=0\,\text{(standard normal)}} = \mu \\
\mathbb{V}[x] &= \int (x-\mu)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\
&= \sigma^2 \underbrace{\int y^2 \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \mathrm{d}y}_{=1\,\text{(standard normal)}} = \sigma^2
\end{aligned}$$

The symbols for location (μ) and squared scale (σ^2) were a hint, but now you know how to show these are equal to the *mean* and *variance*.

Gaussian kurtosis

e.g. Find the kurtosis of the Gaussian

$$\begin{aligned}\sigma_4 &= \int_{-\infty}^{\infty} \left(\frac{x-\mu}{\sigma}\right)^4 \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\ &= \int_{-\infty}^{\infty} y^4 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \mathrm{d}y \\ &= \int_{-\infty}^{\infty} -y^3 \cdot \left(-y\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}\right) \mathrm{d}y \\ &= \underbrace{\left[-y^3 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}\right]_{-\infty}^{\infty}}_{=0} + \int_{-\infty}^{\infty} 3y^2 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \mathrm{d}y \\ &= 3\int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} \mathrm{d}y = 3\end{aligned}$$

Sometimes people define kurtosis against the Gaussian, defining *excess kurtosis* as $\kappa = \sigma_4 - 3$. So a positive excess kurtosis indicates heavy tails of the density and a negative excess kurtosis, the contrary.

Modelling assumptions

e.g. Ashmina is a dietitician. She studies the attention span of a population of school children depending on whether they are given a free nutritional school breakfast or not. The random variable for attention span is t minutes.

She chooses to model t using two Gaussian distributions, one for children given a free nutritional breakfast $\mathcal{N}(t; \mu_1, \sigma_1^2)$, and one for the children who do not receive this breakfast $\mathcal{N}(t; \mu_2, \sigma_2^2)$.

Is this a good model? What conditions should be met for a Gaussian distribution to be a good model?

Other distributions

There are many other distributions, which we shall introduce as we encounter them

e.g.

Exponential	$\frac{1}{Z}e^{-\lambda x}$	$Z = \frac{1}{\lambda}$
Gamma	$\frac{1}{Z}x^{\alpha-1}e^{-\beta x}$	$Z = \frac{\Gamma(\alpha)}{\beta^{\alpha}}$
Beta	$\frac{1}{Z}x^{\alpha-1}(1-x)^{\beta-1}$	$Z = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
Dirichlet	$\frac{1}{Z}\prod_{i=1}^{K} x_i^{\alpha-1}$	$Z = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K}\alpha_i)}$
Student's t	$\frac{1}{Z}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$Z = \frac{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+1}{2})}$

IV: Estimators

Adapted from slides by Dr Herke van Hoof

What is estimation?

We have been looking a bit at this:

$$p(x) \xrightarrow{\text{sampling}} \{x_1, x_2, ..., x_N\}$$

$$p(x) \xrightarrow{\text{expectation}} \{\mu, \sigma^2, ...\}$$

Now we turn our attention to this:

$$\{x_1, x_2, ..., x_N\} \xrightarrow{\text{estimation}} \{\mu, \sigma^2, ...\} \xrightarrow{\text{simple}} p(x)$$

Adapted from slides by Dr Herke van Hoof

What is estimation?

We begin with data $\mathcal{D} = \{x_1, x_2, \dots, x_N\}$, and $x \sim p(x)$. If θ is a property of $p(x)$, what is θ?

We will invent (deterministic) functions $f : \mathcal{D} \to \Theta$, with the desired property

$$f(\mathcal{D}) = \hat{\theta} \simeq \theta$$

estimand value we are trying to guess (θ)

estimator (deterministic) rule to map the dataset to an estimate (f)

estimate result of applying the rule to the dataset ($\hat{\theta}$)

We may have different ways to *estimate* θ, and we want to compare them.

Adapted from slides by Dr Herke van Hoof

What is estimation?

e.g. We have $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$, and we know $x \sim \mathcal{N}(x; \mu, 1^2)$, with unknown location parameter μ. Some estimators of μ are:

- $f(\mathcal{D}) = \overline{x}$ (the sample mean)
- $f(\mathcal{D}) = \tilde{x}$ (the sample median)
- $f(\mathcal{D}) = \frac{\max(\mathcal{D}) - \min(\mathcal{D})}{2}$
- $f(\mathcal{D}) = x_1$ (simply the first point of the sample)
- $f(\mathcal{D}) = \max(\mathcal{D})$
- $f(\mathcal{D}) = \frac{1 + \sum_{i=1}^{N} x_i}{N}$

Some estimators are good, and some are bad. Which of these estimators do you think are reasonable?

Adapted from slides by Dr Herke van Hoof

Properties of estimators

How do we know that the estimator f is appropriate? Here are some properties we would like

First property: Consistency
For a dataset of size N, denote the estimate as $\hat{\theta}_N$. Then if we can show that

$$\lim_{N \to \infty} \hat{\theta}_N = \theta.$$

then our estimator is called *consistent*.

Adapted from slides by Dr Herke van Hoof

Properties of estimators

Second property: Bias

The *bias* of an estimator is the amount it systematically over/undershoots:

$$\text{bias} = \mathbb{E}_{\mathcal{D}}\left[\hat{\theta} - \theta\right]$$

The aim of the game is to find an *unbiased estimator* i.e. bias $= 0$. In the homework you will show sample variance is a biased estimator and fix this.

e.g. Sample mean is an unbiased estimator of the Gaussian location parameter

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\mathbb{E}\left[(x_1 + x_2 + ... + x_n)\right]$$

$$= \frac{1}{N}\left[\mathbb{E}[x_1] + \mathbb{E}[x_2] + ... + \mathbb{E}[x_N]\right] = \frac{1}{N}\underbrace{\left[\mathbb{E}[x] + \mathbb{E}[x] + .. + \mathbb{E}[x]\right]}_{n \text{ times}} = \mathbb{E}[x]$$

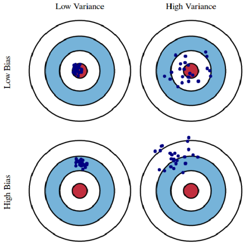
Adapted from slides by Dr Herke van Hoof

Properties of estimators

Third property: Variance
Unbiasedness only measures if errors tend to "average out", it does not tell us how large the errors are. *We prefer small errors*.

We can use variance to measure average error size:

$$\mathbb{V}_{\mathcal{D}}[\hat{\theta}] = \mathbb{E}_{\mathcal{D}}\left[(\hat{\theta} - \mathbb{E}_{\mathcal{D}}[\hat{\theta}])^2\right]$$



(S. Tossato, Fitted Q iteration, 2017)

Arguably, the most important law in ML is **the bias-variance tradeoff**. It states that bias and variance are always in a contest. Minimizing one, raises the other. You cannot have your cake an eat it.

Adapted from slides by Dr Herke van Hoof

Properties of estimators

e.g. Example estimators of the Gaussian location parameter μ:

estimator	consistent?	unbiased?	variance?
$f(\mathcal{D}) = \overline{x}$ (the sample mean)	✓	✓	low
$f(\mathcal{D}) = \tilde{x}$ (the sample median)	✓	✓	low
$f(\mathcal{D}) = \frac{\max(\mathcal{D})-\min(\mathcal{D})}{2}$	✓	✓	high
$f(\mathcal{D}) = x_1$	✗	✓	even higher
$f(\mathcal{D}) = \max(\mathcal{D})$	✗	✗	high
$f(\mathcal{D}) = \frac{1+\sum_{i=1}^{N} x_i}{N}$	✓	✗	low

The biased and inconsistent estimators here are quite silly. In other problems, reasonable estimators can be biased. You'll investigate in the homework.

In the 2nd half of the course, you will see how to obtain reasonable estimators for many distributions.

Adapted from slides by Dr Herke van Hoof

What you should know

- What is a random variable?
- The definition of a distribution's mean and variance, and how to compute it?
- What is a sample mean and variance?
- What are some common distributions and what are their parameters?
- How can we answer simple questions with parametric distributions?
- What is an estimate and what is an estimator?
- What are some important properties of estimators?*

Integration by substitution reminder*

Say we have an integral of the form

$$\int f(g(x))g'(x)\,\mathrm{d}x$$

which is difficult but we know $\int f(y)\,\mathrm{d}y$ is easier. We can *substitute* a new variable $y = g(x)$ so that

$$f(g(x)) = f(y) \qquad \mathrm{d}y = g'(x)\mathrm{d}x.$$

Plugging this back into the integrand we get

$$\int f(g(x))g'(x)\,\mathrm{d}x = \int f(y)\,\mathrm{d}y$$

e.g.

$$\int_0^\infty x \cdot \lambda e^{-\lambda x}\,\mathrm{d}x$$

Substitute $y = \lambda x$, so

$$\begin{aligned} x \cdot \lambda e^{-\lambda x} &= y e^{-y} \\ \mathrm{d}x &= \frac{1}{\lambda}\,\mathrm{d}y \end{aligned}$$

Thus

$$\int_0^\infty x \cdot \lambda e^{-\lambda x}\,\mathrm{d}x = \frac{1}{\lambda}\underbrace{\int_0^\infty y \cdot \lambda e^{-y}\mathrm{d}y}_{=1}$$

Integration by parts reminder*

Notice that from the product rule (of calculus)

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(f(x)g(x)\right) = f'(x)g(x) + f(x)g'(x)$$

$$\underbrace{\int_a^b \frac{\mathrm{d}}{\mathrm{d}x}\left(f(x)g(x)\right)\mathrm{d}x}_{Ⓐ} = \underbrace{\int_a^b f'(x)g(x)\mathrm{d}x}_{Ⓑ} + \underbrace{\int_a^b f(x)g'(x)\mathrm{d}x}_{Ⓒ}$$

$$\implies \underbrace{\int_a^b f(x)g'(x)\mathrm{d}x}_{Ⓒ} = \underbrace{\left[f(x)g(x)\right]_a^b}_{Ⓐ} - \underbrace{\int_a^b f'(x)g(x)\mathrm{d}x}_{Ⓑ}$$

So we can solve integrals of the form

$$\int_a^b f(x)g'(x)\,\mathrm{d}x$$

Integration by parts reminder*

$$\int_a^b f(x)g'(x)\mathrm{d}x = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)\mathrm{d}x$$

e.g.

Solve $\int_0^\infty t\lambda e^{-\lambda t}\,\mathrm{d}t$

$$f(x) = t \qquad g'(x) = \lambda e^{-\lambda t}$$

$$f'(x) = 1 \qquad g(x) = -e^{-\lambda t}$$

$$\int_0^\infty t\lambda e^{-\lambda t}\,\mathrm{d}t = \underbrace{\left[-te^{-\lambda t}\right]_0^\infty}_{=0} + \int_0^\infty e^{-\lambda t}\,\mathrm{d}t = \left[-\frac{1}{\lambda}e^{-\lambda t}\right]_0^\infty = \frac{1}{\lambda}$$