Week 2a:

Probability theory for machine learning

G6061: Fundamentals of Machine Learning [23/24]

Dr. Johanna Senk

UNIVERSITY OF SUSSEX

Recap of previous lecture

Classification and evaluation of classifiers

- Binary classification
  - Confusion matrix
  - Metrics:
    Accuracy, Error
    Sensitivity, Specificity
    Precision, Recall

	Predicted +	Predicted -	
Actual +	30	20	50
Actual -	10	90	100
	40	110	150

- Multi-class classification
  by combining several binary classifiers
  - One-versus-rest (OVR) strategy
  - One-versus-one (OVO) strategy

UNIVERSITY OF SUSSEX

Probability and machine learning

- ML all about reducing our uncertainty.
- Good applications of ML account for uncertainty before and after applying ML.
- How to understand uncertainty?

PROBABILITY AND STATISTICS!

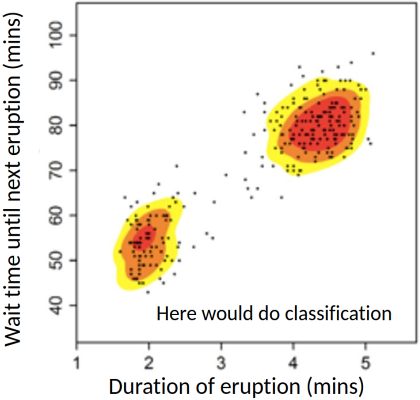
Why is probability density estimation useful?

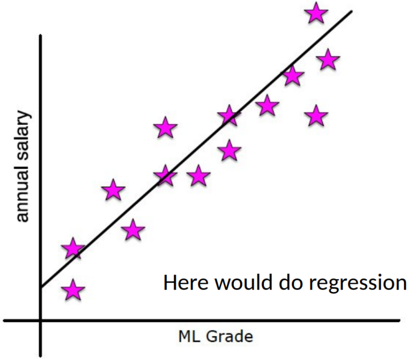
- **For designing your ML method!**
  It's a lot easier to classify data if you have the underlying distributions.
- Build up a probability distribution from previous instances.
- Understand how distributions from two or more classes overlap, to inform choice of machine learning algorithm.

- Probability density $\approx$ probability distribution.
  For variables that can vary continously, use a density to define more likely and less likely regions where data samples might lie.

Examples

Old Faithful Geyser,
Yellowstone National Park, Wyoming





Important points for interpreting ML results

- The accuracy of a ML algorithm may change on new data.
- Unless you've tested your algorithm on an enormous dataset, your estimate of the accuracy might itself not be that accurate!
  - I have two classifiers, one got 86% accuracy, one got 90% accuracy, do I know for sure which one is better?

Overview

- Why a good understanding of probability is important in ML.

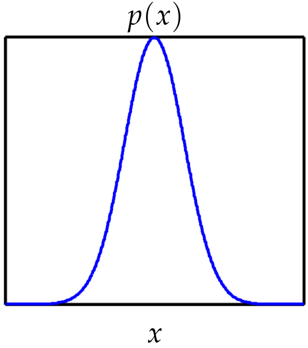
Probability theory

- Probability density functions
- Properties / parameters of probability distributions
- Multivariate probability distributions
- Uniform distribution and Gaussian distribution (aka normal distribution)

UNIVERSITY OF SUSSEX

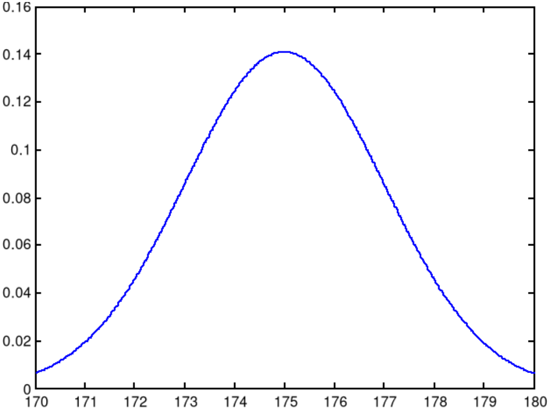
Probability theory

X	Random variable (r.v.)
x	Outcome of X
$X\|Y = y$	Conditional r.v. for X given that $Y = y$
$P(X = x)$	Probability that $X = x$ (discrete variable)
$p(x)$	Probability density (continuous variable)



Probability density functions

Very often encounter continuous variables in ML, and their distributions are given by:

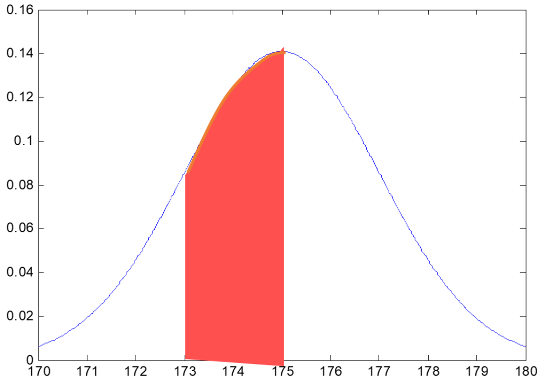


Height (cm)

The function known as a **probability density function (pdf)**.

The higher the probability density at x and around x, the more likely that the value of a datapoint will be close to x.

Probability density functions



Height (cm)

- Probability:
  Area under the curve
  $P(173 \leq X \leq 175) = \int_{173}^{175} p(x)\mathrm{d}x$

- Normalisation:
  Area under whole curve must sum to 1
  $1 = \int_{-\infty}^{\infty} p(x)\mathrm{d}x$

Properties / parameters of distributions

- Distributions and pdfs are often described by parameters, commonly mean and variance (or mean and standard deviation).
- The **mean** is the usual average, also known as the expected value $E(X)$ or $\langle X \rangle$:

$$\langle X \rangle = \sum_x x P(X = x) \qquad \text{or} \qquad \int_{-\infty}^{\infty} x p(x) \mathrm{d}x$$

- Compare this with the **sample mean** (= estimate of the distribution), written as $\overline{x}$:

$$\overline{x} = \frac{1}{n} \sum_i x_i$$

For the probability distribution version, the sum of all the P's is 1, so the denominator is just 1.

Properties / parameters of distributions

- Distributions and pdfs are often described by parameters, commonly mean and variance (or mean and standard deviation).
- The **mean** is the usual average, also known as the expected value $E(X)$ or $\langle X \rangle$:

$$\langle X \rangle = \sum_x x P(X = x) \qquad \text{or} \qquad \int_{-\infty}^{\infty} x p(x) \mathrm{d}x$$

- The **variance** governs the spread of the data. It is given by the expected square of the deviation from the mean:

$$Var(X) = \langle (X - \langle X \rangle)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2$$

  In other words, the mean value of the square of the distance from the mean.
- The **standard deviation** is the square root of the variance, and is similar to an average distance from the mean.

UNIVERSITY OF SUSSEX

Properties / parameters of distributions

	Mean $\langle X \rangle, \mathbb{E}(X), \mu$	**Variance** $Var(X), \mathbb{V}(X), \sigma^2$
Discrete (distribution)	$\langle X \rangle = \sum_x x P(X = x)$	$Var(X) = \sum_x (x - \langle X \rangle)^2 P(X = x)$
Continuous (pdf)	$\langle X \rangle = \int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x$	$Var(X) = \int_{-\infty}^{\infty} (x - \langle X \rangle)^2 p(x) \, \mathrm{d}x$
Sample	$\overline{x} = \frac{1}{n} \sum_i x_i$	$Var = \frac{1}{n} \sum_i (x_i - \overline{x})^2$

- **Standard deviation**: $\sigma_X = \sqrt{Var(X)}$

UNIVERSITY OF SUSSEX

Example: Fair 6-sided dice

$$P(X=1) = P(X=2) = P(X=3) = P(X=4) = P(X=5) = P(X=6) = 1/6$$

- **Mean**:
  $\langle X \rangle = \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = 3.5$

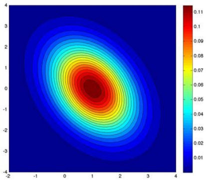
- **Variance**:
  $Var(X) = \langle (X - \langle X \rangle)^2 \rangle = \frac{1}{6}\left((\frac{5}{2})^2 + (\frac{3}{2})^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2 + (\frac{3}{2})^2 + (\frac{5}{2})^2 \right) = \frac{35}{12}$
- **Standard deviation**:
  $\sigma_X = \sqrt{Var(X)} \approx 1.71$

- C.f. mean distance from mean is 1.5
  Standard deviation easier to work with mathematically than mean distance from mean.

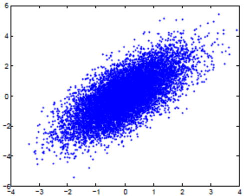
UNIVERSITY OF SUSSEX

Multivariate probability distributions

- By multivariate we mean multi-dimensional, i.e., 2 or more random variables.



Example density



Example sample

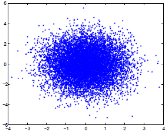
- Multi-dimensional integral of probability density function to get probability a sample will lie within a certain region.

- E.g., 2 variables:
  $P\left((X,Y) \text{ in region}\right) = \int_{(x,y) \text{in region}} p(x,y)\,\mathrm{d}x\mathrm{d}y$
- Or for discrete variables:
  $P\left((X,Y) \text{ in region}\right) = \sum_{(x,y) \text{in region}} P(X=x, Y=y)$

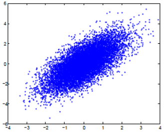
Covariance and correlation

- **Covariance**:

$$Cov(X, Y) = \langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle = \int (x - \langle X \rangle)(y - \langle Y \rangle) p(x, y) \, \mathrm{d}x\mathrm{d}y$$



Example zero cov.



Example positive cov.

- **Correlation** is a normalised covariance:

$$Corr(X, Y) = \frac{\langle (X - \langle X \rangle)(Y - \langle Y \rangle) \rangle}{\sigma_X \sigma_Y} \qquad -1 \leq Corr(X, Y) \leq 1$$

- In sample, estimate it with:

$$Cov = \frac{1}{n} \sum_i (x_i - \overline{x})(y_i - \overline{y})$$

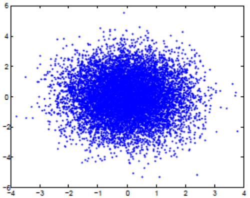
UNIVERSITY OF SUSSEX

Independence

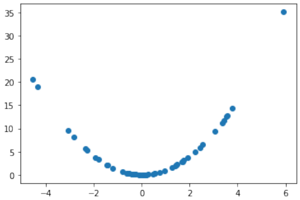
- Two variables X and Y are independent if for each x and y:

$$p(x,y) = p(x)p(y) \quad \text{or equivalently} \quad p(x|y) = p(x)$$

- Independent variables have zero covariance (and correlation)
- But zero covariance (and correlation) does not imply independence!
- Both these examples have zero covariance (and correlation):



Independent



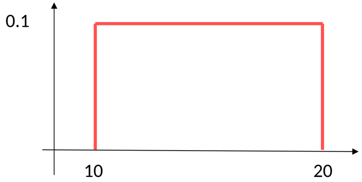
Not independent

UNIVERSITY OF SUSSEX

Common distributions

- Here will meet the two most common distributions:
  1. uniform
  2. Gaussian, or normal
- Other distributions include binomial, multinomial, Poisson etc. (You can look these up on Mathworld or Wikipedia.)

The **uniform distribution** has the same probability for each point. Thus probability is governed by the range of the data R and pdf $p(x) = 1/R$, which equals 0.1 in the example below:

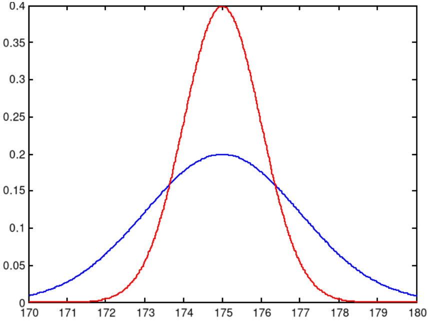


Gaussian distribution

- pdf:

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

- It is centred on μ and width (and height) are governed by σ^2.

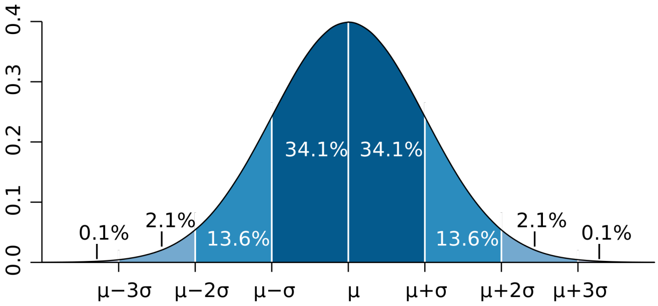


Height (cm)

- red has $\sigma^2 = 1\ (\sigma = 1)$
- blue has $\sigma^2 = 4\ (\sigma = 2)$

UNIVERSITY OF SUSSEX

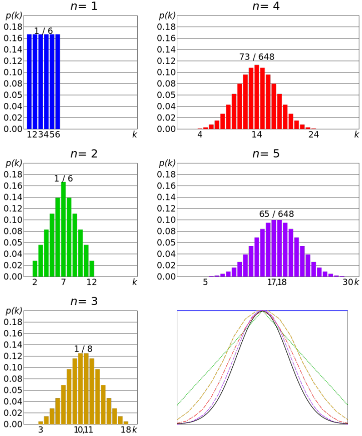
Gaussian distribution



Wikipedia (CC-BY-SA 3.0)

- 68% chance of being within 1 sd of mean
- 95% chance of being within 2 sd of mean
- 99.7% chance of being with 3 sd of mean
- 99.99994% chance of being with 5 sd of mean

Central limit theorem



Wikipedia (CC-BY-SA 3.0)

- The distribution of the sum (or the mean) of n i.i.d. (independent identically distributed) random variables becomes increasingly Gaussian as n grows.
- Sum:

$$X = \sum_{i=1}^{n} U_i$$

- Mean:

$$X = \frac{1}{n} \sum_{i=1}^{n} U_i$$

- Example: Rolling a fair dice n times.

UNIVERSITY OF SUSSEX

Multivariate Gaussian distribution

- Single normal random variable with mean μ and standard deviation σ:

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

- Single normal random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

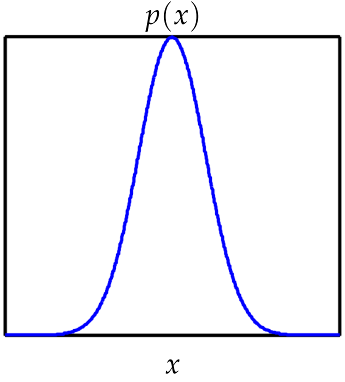
$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

$$\Sigma = \begin{pmatrix} var(x_1) & cov(x_1,x_2) & \ldots & cov(x_1,x_D) \\ cov(x_1,x_2) & var(x_2) & \ldots & cov(x_2,x_D) \\ \ldots & \ldots & \ldots & \ldots \\ cov(x_1,x_D) & cov(x_2,x_D) & \ldots & var(x_D) \end{pmatrix}$$

UNIVERSITY OF SUSSEX

Summary and outlook

- Probability density functions:
  $1 = \int_{-\infty}^{\infty} p(x)\mathrm{d}x$
- Properties / parameters of probability distributions:
  mean, variance, standard deviation
- Multivariate probability distributions:
  covariance, correlation, independence
- Uniform distribution and Gaussian distribution
  (aka normal distribution):
  central limit theorem



- Next lecture:
  Bayes' theorem, (non-)parametric probability
  density estimation