Week 2b:

Probability theory for machine learning

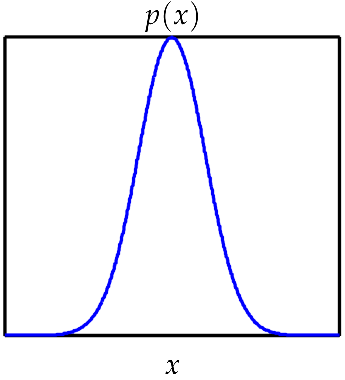
G6061: Fundamentals of Machine Learning [23/24]

Dr. Johanna Senk

UNIVERSITY
OF SUSSEX

Recap of previous lecture

- 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



UNIVERSITY OF SUSSEX

Warm-up: Heads or tails?

- I want to know how to test whether a coin is biased when it comes to landing on heads or tails. To do this, I'll investigate the probability distribution for the proportion of throws that come up heads for a fair coin.
- Let's analyse the distribution of this if I just throw the coin twice. In this case:

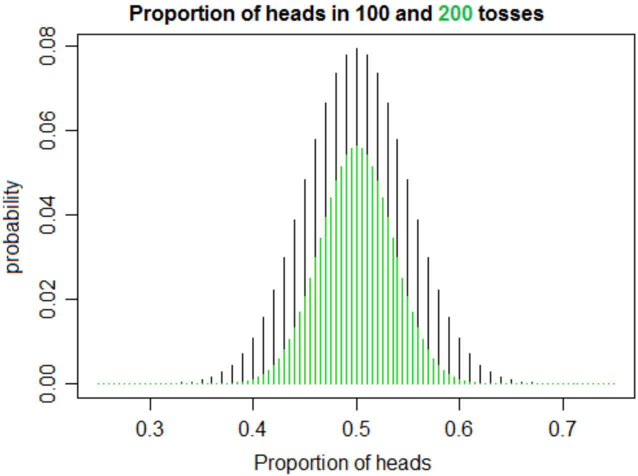
$$\begin{array}{ll} P(X=0)=1/4 & \text{Two tails} \\ P(X=1/2)=1/2 & \text{Tails then heads, or heads then tails} \\ P(X=1)=1/4 & \text{Two heads} \end{array}$$

- Mean? $E(X)=\langle X\rangle=\frac{1}{4}\cdot 0+\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{4}\cdot 1=\frac{1}{2}$
- Variance? $Var(X)=\langle(X-\langle X\rangle)^2\rangle=\frac{1}{4}\cdot\left(\frac{1}{2}\right)^2+\frac{1}{2}\cdot 0+\frac{1}{4}\cdot\left(\frac{1}{2}\right)^2=\frac{1}{8}$
- Standard deviation? $\sigma_X=\sqrt{Var(X)}\approx 0.35$

UNIVERSITY OF SUSSEX

Is the coin fair?

We need a lot more than two tosses to test if a coin is fair!



Keep this in mind when assessing the accuracy of an ML algorithm!

Overview

Probability distributions are important in ML:

- To characterize your data and inform your choice/design of algorithm.
- To interpret ML results.

Today:

- Application of **Bayes' theorem** to interpret results
- **Probability density estimation**
  - Non-parametric approach (histograms, kernel density estimation)
  - Parametric approach

UNIVERSITY OF SUSSEX

Bayes' theorem

- Very (very) useful theorem:

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

- Follows from **conditional probability** and **joint probability** relation:

$$\begin{aligned} P(A, B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

The joint probability of two events equals the probability of event A times the probability of event B given event A.

Bayes' theorem and classifiers

If the classifier says 1, what is the probability that the class is actually 1?

This is the sensitivity of the classifier for detecting class 1

$$P(\text{Class}=1 \mid \text{Classifier says 1}) = \frac{P(\text{Classifier says 1} \mid \text{Class}=1)\, P(\text{Class}=1)}{P(\text{Classifier says 1})}$$

Our prior expectations

This denominator term we don't know

Want to know this

UNIVERSITY OF SUSSEX

Bayes' theorem and classifiers

If the classifier says 1, what is the probability that the class is actually 1?

$$P(\text{Class} = 1|\text{Classifier says } 1) = \frac{P(\text{Classifier says } 1|\text{Class} = 1)P(\text{Class} = 1)}{P(\text{Classifier says } 1)}$$

$$P(\text{Class} = 0|\text{Classifier says } 1) = \frac{P(\text{Classifier says } 1|\text{Class} = 0)P(\text{Class} = 0)}{P(\text{Classifier says } 1)}$$

Compute the **odds ratio** for class 1 vs. class 0, assuming our **prior expectation** is correct:

$$\frac{P(\text{Class} = 1|\text{Classifier says } 1)}{P(\text{Class} = 0|\text{Classifier says } 1)} = \frac{P(\text{Classifier says } 1|\text{Class} = 1)P(\text{Class} = 1)}{P(\text{Classifier says } 1|\text{Class} = 0)P(\text{Class} = 0)}$$

UNIVERSITY OF SUSSEX

Bayes' example: COVID test

- Suppose there's a new COVID test which picks up COVID early, at the first hint of symptoms. The probability of testing positive given you actually have COVID (= **sensitivity**) is $P(\text{T=1}|\text{COVID=1}) = 0.99$.
- However, the **specificity** is not as good as the sensitivity, it has a 10% false positive rate: $P(\text{T=1}|\text{COVID=0}) = 0.1$
- Are these tests useful? Depends on your **prior**: $P(\text{COVID=1}) = ?$
- Odds ratio to compute:

$$\frac{P(\text{COVID} = 1|\text{T} = 1)}{P(\text{COVID} = 0|\text{T} = 1)} = \frac{P(\text{T} = 1|\text{COVID} = 1)P(\text{COVID} = 1)}{P(\text{T} = 1|\text{COVID} = 0)P(\text{COVID} = 0)}$$

UNIVERSITY OF SUSSEX

Bayes' example: COVID test

- **Case 1**: Tonnes of COVID around: prior $P(\text{COVID=1}) = 0.5$

$$\frac{P(\text{COVID} = 1|\text{T} = 1)}{P(\text{COVID} = 0|\text{T} = 1)} = \frac{P(\text{T} = 1|\text{COVID} = 1)P(\text{COVID} = 1)}{P(\text{T} = 1|\text{COVID} = 0)P(\text{COVID} = 0)} = \frac{0.99 \cdot 0.5}{0.1 \cdot 0.5} = \frac{0.495}{0.05} = 9.9$$

10 times more likely than not to have COVID.

- **Case 1**: Not much COVID around: prior $P(\text{COVID=1}) = 0.05$

$$\frac{P(\text{COVID} = 1|\text{T} = 1)}{P(\text{COVID} = 0|\text{T} = 1)} = \frac{P(\text{T} = 1|\text{COVID} = 1)P(\text{COVID} = 1)}{P(\text{T} = 1|\text{COVID} = 0)P(\text{COVID} = 0)} = \frac{0.99 \cdot 0.05}{0.1 \cdot 0.95} = \frac{0.0495}{0.095} = 0.52$$

Roughly half as likely to have COVID than not have COVID,
i.e., about 1 in 3 chance of having COVID.

Probability density estimation

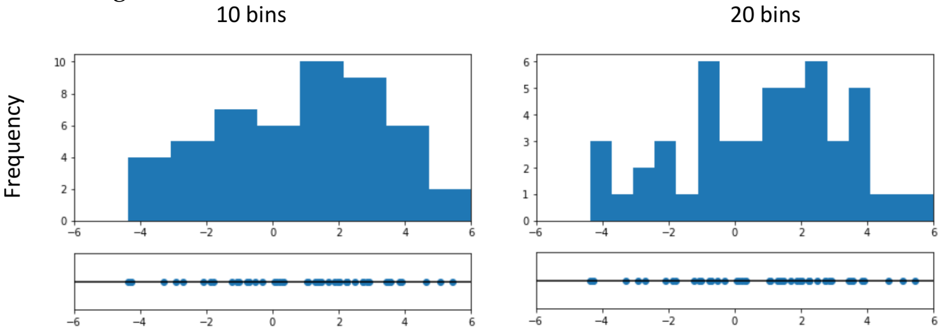
- **Non-parametric estimation (histograms, kernel density estimation)**:
  No assumptions about the form of the probability density function, it is determined entirely from the data.

- **Parametric estimation**:
  Assumes a specific kind of distribution, e.g., Gaussian (normal).
  Parameters of the distribution are optimized to fit the data (usually mean, standard deviation, plus covariances if multi-dimensional).

UNIVERSITY OF SUSSEX

Non parametric method: Histograms

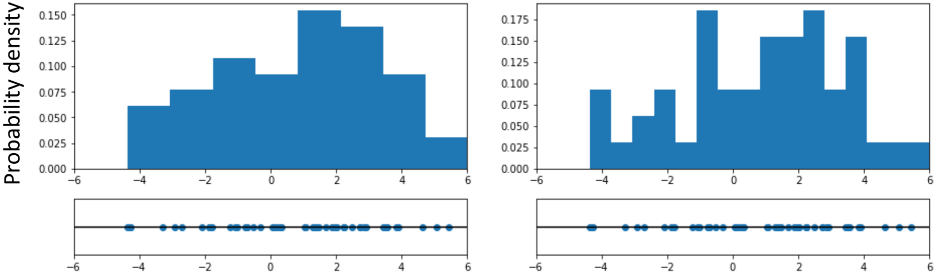
- Divide range of data into a certain number of **bins** and plot number of data points that fall in each bin.
- 50 data points, drawn from a normal distribution, but would you know from these histograms that the distribution is normal?



UNIVERSITY OF SUSSEX

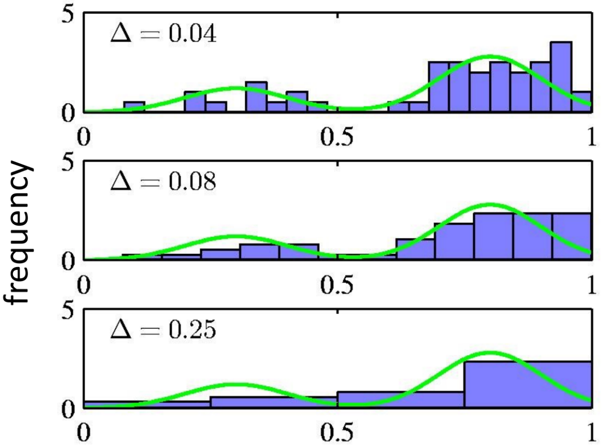
Non parametric method: Histograms

- **Normalise** the bars so that their heights represent **probability density** (i.e., rescale y-axis).
- Proportion of data that lie in the bin is given by: (height of bar) x (width of bar) (corresponding to probabilities being determined by areas under a probability density curve).
- Sum of areas of all bars = 1.

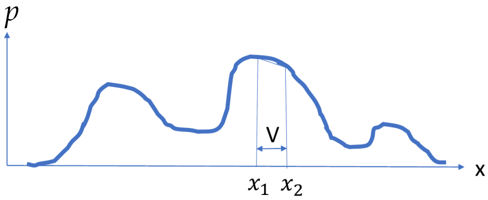


Non parametric method: Histograms

- Choice of **bin width**, Δ, can impact conclusions, so should be considered carefully.
- How many classes do we have? It should be two here, but histogram may or may not show that.
- Horizontal alignment of bars also important. Are the bars centred on the left or right bin edge or between the bin edges?



Using a histrogram



- p is the probability density
- N is the sample size
- k is the number of points in small range V

$P(X \text{ lies in small range of length } V) = k/N$ Good approx. if N and k tend to be large.
$P(X \text{ lies in small range of length } V) = pV$ Good approx. if V is small.

$$pV = k/N \quad \text{so estimate} \quad p = \frac{k}{NV}$$

UNIVERSITY OF SUSSEX

Kernel density estimation

- Instead of density at x just being number of points within a fixed small distance of x, do a weighting, so data points very close to x contribute a lot, and points further from x contribute little.

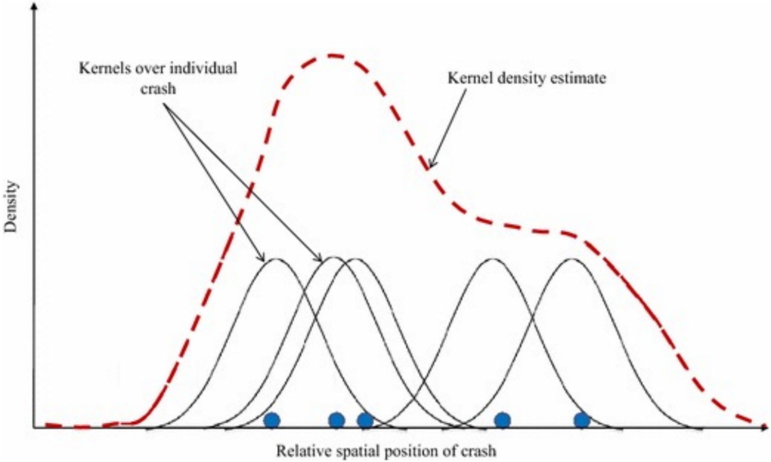
$$p = \frac{k}{NV} \quad \rightarrow \quad p(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{V} K\left(\frac{x - x_i}{V}\right)$$

- A Gaussian function can be used for the kernel K in which case V is its standard deviation.

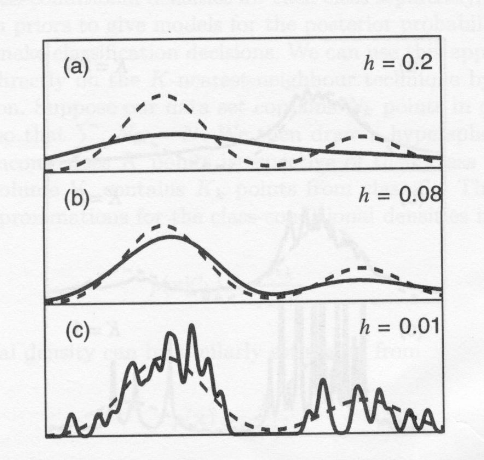
UNIVERSITY OF SUSSEX

Example

Distribution of car crashes along a road



Kernel density estimation



It is possible to use different forms of kernel to get smoother continuous estimates for x.

V is critical:

- too small $\rightarrow$ spiky pdf
- too big $\rightarrow$ over-smoothed

Parametric density estimation

- Assume a particular kind of distribution and then make your best guess of the parameters. For the example of the **Gaussian (normal) distribution**, the probability density function (pdf) is:

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

- The parameters to find in this case are
  the mean $\mu = E(X) = \langle X \rangle$
  and the variance $\sigma^2 = E(X-\mu)^2 = \langle (X-\mu)^2 \rangle$

UNIVERSITY OF SUSSEX

Parametric density estimation

Naïve approach:

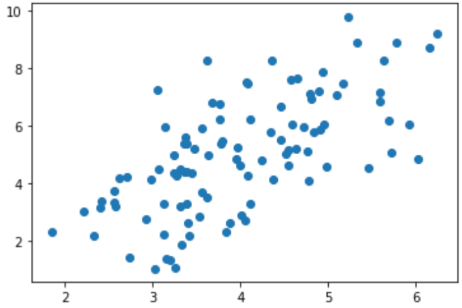
- The most obvious way of estimating the mean and variance is simply to take the mean and variance of the sample.

- Guess for the mean μ:
  $\overline{x} = \frac{1}{n}\sum_i x_i$
- Guess for the variance σ^2:
  $Var = \frac{1}{n}\sum_i (x_i - \overline{x})^2$

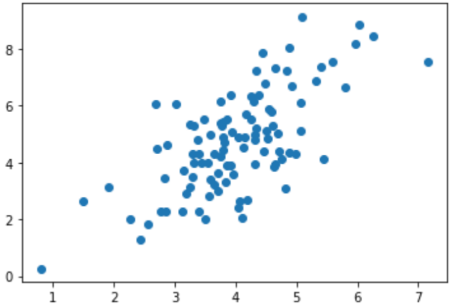
UNIVERSITY OF SUSSEX

Example

- For a given dataset, you don't know how accurate your estimates are. Consider the following two samples, both with true mean $(4, 5)$:



Mean=(4.00,5.02)



Mean=(4.01,4.74)

UNIVERSITY OF SUSSEX

Confidence intervals

- You never know for sure how good your estimate of the mean and standard deviation are.
- But we can compute the **standard error**, which is roughly what the standard deviation of the estimate of the mean would be if we repeated the experiment many times:

$$\text{Standard error} = \frac{\text{Standard deviation (data)}}{\sqrt{\text{number of data points}}}$$

- A very rough "rule-of-thumb" is that the true mean is unlikely to be more than two standard errors away from your estimate.

UNIVERSITY OF SUSSEX

Different methods

There are more sophisticated methods for parametric density estimation to be aware of:

- **Maximum likelihood estimation**:
  Choose the parameters that maximise the overall probability density function for the n data points that you have.

- **Bayesian inference**:
  Parameters θ described by a probability distribution. Initially set to prior distribution and converted to posterior $P(\theta|X)$ through Bayes' theorem once data is observed.

UNIVERSITY OF SUSSEX

Maximum likelihood estimation

- pdf of normal (Gaussian) distribution:

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

- **Likelihood** $\mathcal{L}(\mu, \sigma^2)$ = pdf for n i.i.d. normal random variables:

$$p(x_1, \dots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

- Optimisation problem: We have to find the mean and standard deviation that maximise the joint probability density.
- Values which maximize the likelihood will also maximize its logarithm, the **log-likelihood** $\log\Big(\mathcal{L}(\mu, \sigma^2)\Big)$.
- For the normal distribution, the most likely mean is the mean of the data (= sample mean) and the most likely standard deviation is the standard deviation of the data. For other distributions it can get more complicated.

UNIVERSITY OF SUSSEX

Multivariate normal

- Suppose we have a multivariate normal, say overall levels of red R, green G and blue B in an image that is part of a large dataset.

- We need to find the 3 means and the 3 variances.
  What other parameters are there?

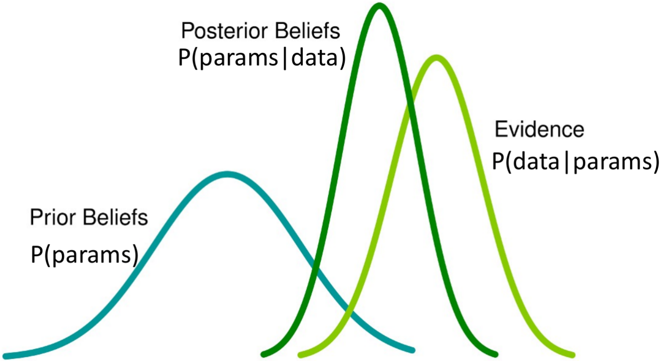
Multivariate normal

- Suppose we have a multivariate normal, say overall levels of red R, green G and blue B in an image that is part of a large dataset.
- We need to find the 3 means and the 3 variances.
  What other parameters are there?
- Three covariances! $Cov(R, G)$, $Cov(R, B)$, $Cov(G, B)$.

UNIVERSITY OF SUSSEX

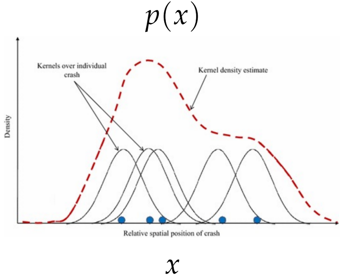
Bayesian inference

- $P(\text{params}|\text{data})$ proportional to $P(\text{data}|\text{params}) \times P(\text{params})$
- Given the data, get a new likely range for the parameters.



Summary and outlook

- What have you learned about estimating probability density functions?
  - Can do it **non-parametrically**:
    - Use histograms to estimate the density.
    - Kernel density estimation is like a smoothed-out histogram, where each data point contributes to the density estimate in a region around it.
  - Can do it **parametrically**:
    - By assuming the form of a distribution (often Gaussian) and
    - Finding the best fit parameters - usually mean, standard deviation (plus covariances if multi-dimensional).



- Next lecture: Linear regression

UNIVERSITY OF SUSSEX