COMS21202: Symbols, Patterns and Signals
Probabilistic Data Models

[based on Dima Damen lecture notes]

Rui Ponte Costa

rui.costa@bristol.ac.uk

Department of Computer Science, University of Bristol
Bristol BS8 1UB, UK

February 6, 2020

Data Modelling

- Deterministic models do not explicitly model uncertainties or ‘randomness’ in data
- Inference variability from the data is not included
- In many tasks, we benefit from modelling uncertainty
- This is explicit in **Probabilistic Models**

Back to Fish - Discrete case

Discrete variable:

Example

A fisherman returns with the daily catch of fish. If we select a fish at random from the hold, what species will it be?

$$fish \in \{salmon, seabass, cod, ...\}$$

- A deterministic model would give **one** value, the most likely
- A probabilistic model quantifies the chance/probability of the selected fish being one of the possible species.
- Model the probability $P(x_i = q_i)$ where $q_i \in \{salmon, seabass, cod, \cdots\}$

Back to Fish - Continuous case

Continuous variable:

Example

Predict the weight of fish from its length

Let us assume that we think the weight of fish is directly proportional to its length, i.e. $\textit{weight} = b \times \textit{length} + a$ (linear regression)

A **probabilistic approach** would model weight as a **random variable** and hypothesize that

$$\textit{weight} = b \times \textit{length} + a + \epsilon$$

where ϵ is a random variable, with mean usually close to zero

Back to Fish - Continuous case

$$weight = b \times length + a + \epsilon$$

- Modelled using a probability distribution for ϵ,
  - by a uniform distribution
  - by a normal distribution
  - ...
- In the next slides, we will simplify things by setting weight = 0 when length = 0
- As a consequence, the y-intercept can be set to zero (i.e. a=0), and

$$weight = b \times length + \epsilon$$

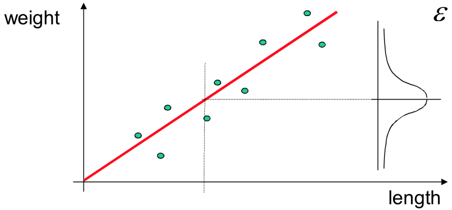
Back to Fish - Probabilistic

$$weight = b \times length + \epsilon$$

We can assume, for example, that ϵ is given by $\mathcal{N}(0, \sigma^2)$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\epsilon^2}{2\sigma^2}}$$

This model has two parameters: the slope b and uncertainty σ (with $\mu = 0$)



Maximum Likelihood Estimation

- Similar to building deterministic models, probabilistic model parameters need to be tuned/trained
- **Maximum-likelihood estimation (MLE)** is a method of estimating the parameters of a probabilistic model.

- Assume $\boldsymbol{\theta}$ is a vector of all parameters of the probabilistic model. (e.g. $\boldsymbol{\theta} = \{b, \sigma\}$).
- **MLE** is an extremum estimator[1] obtained by maximising an objective function of $\boldsymbol{\theta}$

[1] "Extremum estimators are a wide class of estimators for parametric models that are calculated through maximization (or minimization) of a certain objective function, which depends on the data." wikipedia.org

Maximum Likelihood Estimation

Definition

Assume $f(\theta)$ is an objective function to be optimised (e.g. maximised), the *arg max* corresponds to the value of θ that attains the maximum value of the objective function *f*

$$\hat{\theta} = \textit{arg max}_\theta\, f(\theta)$$

- Tuning the parameter is then equal to finding the maximum argument *arg max*

Maximum Likelihood Estimation - General

- Maximum Likelihood Estimation (MLE) is a common method for solving such problems

$$
\begin{aligned}
\theta_{MLE} &= \textit{arg max}_\theta \, p(D|\theta) \\
&= \textit{arg max}_\theta \, \ln p(D|\theta) \\
&= \textit{arg min}_\theta \, -\ln p(D|\theta)
\end{aligned}
$$

MLE Recipe

1. Determine θ, D and expression for likelihood $p(D|\theta)$
2. Take the natural logarithm of the likelihood
3. Take the derivative of $\ln p(D|\theta)$ w.r.t. θ. If θ is a multi-dimensional vector, take partial derivatives
4. Set derivative(s) to 0 and solve for θ

MLE: 1. Define likelihood

Given a set of N data points - x_i is length and y_i is weight in our *fishy* example

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

- The probabilistic approach would:
  - derive expression for conditional probability of observing data D given parameters $\boldsymbol{\theta} = \{b, \sigma\}$

$$p(D|\theta)$$

MLE: 1. Define likelihood

Given a set of N data points

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

Assume that observations are independent - a common assumption often referred to as **i.i.d. independent and identically distributed** - then :

$$p(D|\theta) = \prod_{i=1}^{N} p(y_i|x_i, \theta)$$

Given $y_i = b\,x_i + \epsilon$, and ϵ is $\mathcal{N}(0, \sigma^2)$, then

$$p(y_i|x_i, \theta) \sim \mathcal{N}(bx_i, \sigma^2)$$

For a large sample:

- The average of y_i value will be $b\,x_i$
- The 'spread' or variance will be the same as for ϵ, defined by σ^2

MLE: 1. Define likelihood

The conditional probability (for all data) is thus formulated as

$$
\begin{aligned}
p(D|\theta) &= \prod_{i=1}^{N} p(y_i|x_i,\theta) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_i - bx_i)^2}{\sigma^2}}
\end{aligned}
$$

MLE: 2. Take natural logarithm

We will focus on parameter b for the next steps,

$$
\begin{aligned}
b_{ML} &= arg\ max_b\, p(D|\theta) \\
&= arg\ max_b \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_i - bx_i)^2}{\sigma^2}} \\
&= arg\ max_b \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_i - bx_i)^2}{\sigma^2}}\right) && \text{(use ln trick)} \\
&= arg\ max_b \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_i - bx_i)^2}{\sigma^2}}\right) && \text{(ln prop. 1)} \\
&= arg\ max_b \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_i - bx_i)^2}{2\sigma^2} && \text{(ln prop. 2)} \\
&= arg\ max_b \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(y_i - bx_i)^2 && \text{(discard no-b terms } \ln\frac{1}{\sqrt{2\pi}\sigma}\text{)} \\
&= arg\ min_b \sum_{i=1}^{N} \frac{1}{2\sigma^2}(y_i - bx_i)^2 && \text{(switch to minimisation form.)}
\end{aligned}
$$

Data Modelling - Deterministic vs Probabilistic

- Deterministic Least Squares [Lecture 3]:

$$b_{LS} = arg\, min_b\, R(b, a = 0) = arg\, min_b \sum_i (y_i - b\, x_i)^2$$

- Probabilistic Maximum Likelihood:

$$b_{ML} = arg\, min_b \sum_i \frac{1}{2\sigma^2} (y_i - b\, x_i)^2$$

- probabilistic model explicit considers uncertainty, σ^2.

MLE: 3. Take derivatives and 4. Find solution

To simplify the calculations here we assume $\sigma = 1$,

$$b_{ML} = arg\, min_b \sum_i \frac{1}{2}(y_i - b\, x_i)^2$$

To find the minimum, calculate the derivative

$$\frac{d}{db}\sum_i \frac{1}{2}(y_i - bx_i)^2 = -\sum_i x_i(y_i - bx_i)$$

and equate it to zero

$$-\sum_i x_i(y_i - b_{ML}x_i) = 0$$

$$\sum_i x_i y_i - b_{ML}\sum_i x_i^2 = 0$$

$$b_{ML} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

MLE: summary

Example: normal distribution (parameter *b*)

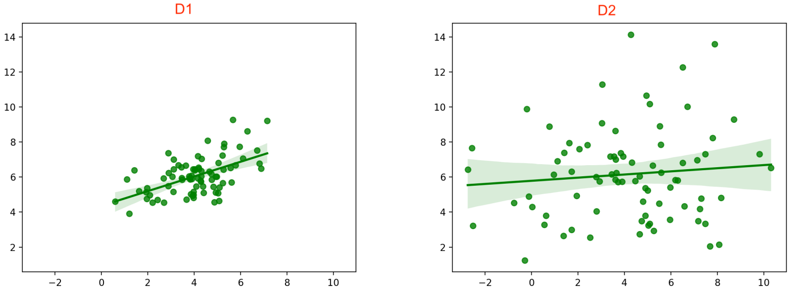
1. Determine θ, D and expression for likelihood $p(D|\theta)$
   $p(D|b) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(y_i - bx_i)^2}{\sigma^2}}$
2. Take the natural logarithm of the likelihood
   $b_{ML} = arg\, min_b \sum_i \frac{1}{2\sigma^2}(y_i - b\,x_i)^2$
3. Take the derivative of $\ln p(D|\theta)$ w.r.t. θ. If θ is a multi-dimensional vector, take partial derivatives [2]
   $\frac{d}{db}\sum_i \frac{1}{2}(y_i - bx_i)^2 = -\sum_i x_i(y_i - bx_i)$
4. Set derivative(s) to 0 and solve for b
   $b_{ML} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$

[2] For simplicity we set $\sigma = 1$

Rui Ponte Costa
rui.costa@bristol.ac.uk

Data Modelling - Deterministic vs Probabilistic

- **Probabilistic Models** can tell us more
- We could use the same MLE recipe to find σ_{ML}. This would tell us how uncertainty our model is about the data D.
- For example: if we apply this method to two datasets (D_1 and D_2) what would the parameters $\theta = \{b, \sigma\}$ be?



$b_{ML}^{D_1} > b_{ML}^{D_2}$ [slope] and $\sigma_{ML}^{D_1} < \sigma_{ML}^{D_2}$ [uncertainty[3]]

[3] The uncertainty (σ) is represented by the light green bar in the plots. Test it your self.

Probabilistic Model - Ex2

Example

Given a coin, you were assigned the task of figuring out whether the coin will land on its head or tails. You were asked to build a probabilistic model (i.e. with confidence)

- **Data:** head/tail binary attempts (of size N)
- **Model:** Binomial distribution
- **Model Parameters:** head probability α

Probabilistic Model - Ex2

Definition

The **binomial distribution** gives the probability distribution for a discrete variable to obtain exactly D successes out of N trials, where the probability of the success is α and the probability of failure is $(1-\alpha)$ and $0 \leq \alpha \leq 1$

The binomial distribution probability density function is given by

$$P(D|N) = \binom{N}{D}\alpha^{D}(1-\alpha)^{N-D}$$
$$= \frac{N!}{D!(N-D)!}\alpha^{D}(1-\alpha)^{N-D}$$

Probabilistic Model - Ex2

Accordingly, using the binomial probability distribution where D is the number of heads in N coin tosses and θ is the probability of getting heads in a single toss,

$$P(D|\theta) = \binom{N}{D}\theta^D(1-\theta)^{N-D}$$

Maximum Likelihood Estimation (MLE) would then be looking for

$$\theta_{ML} = arg\, max_\theta\, p(D|\theta)$$

Probabilistic Model - Ex2

- Take the natural logarithm

$$P(D|\theta) = \binom{N}{D}\theta^D(1-\theta)^{N-D}$$

$$\ln P(D|\theta) = \ln\binom{N}{D} + D\ln\theta + (N-D)\ln(1-\theta)$$

- Take the derivative w.r.t θ

$$\frac{d}{d\theta}\ln P(D|\theta) = D\frac{1}{\theta} + (N-D)\frac{1}{1-\theta}(-1)$$

$$= \frac{D}{\theta} - \frac{N-D}{1-\theta}$$

Probabilistic Model - Ex2

- Set the derivative to 0 and solve for θ

$$\frac{D}{\theta_{ML}} - \frac{N-D}{1-\theta_{ML}} = 0$$

$$\frac{D(1-\theta_{ML}) - (N-D)\theta_{ML}}{\theta_{ML}(1-\theta_{ML})} = 0$$

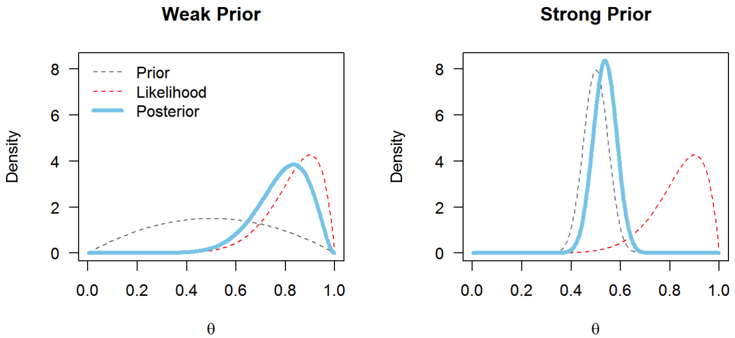
$$D - N\theta_{ML} = 0$$

$$\theta_{ML} = \frac{D}{N}$$

- In conclusion, the probability of *heads* is the relative frequency (D over the number of samples N).

Probabilistic Model - Using prior information

- MLE ignores any prior knowledge we may have about θ
- If we have prior knowledge about values what θ is likely to have, then we can use Bayesian inference, which combines prior and likelihood probabilities as $p(\theta|D) = p(D|\theta)p(\theta)/Z$, where $p(\theta|D)$ is the posterior, $p(D|\theta)$ the likelihood of the data, $p(\theta)$ the prior over the parameters and Z is the normalization term (p(D)).



https://jimgrange.wordpress.com/2016/01/18/pesky-priors/

Probabilistic Model - Using prior information

- The MLE method can be expanded to consider prior information as

$$\theta_{ML} = arg\, max_{\theta}\, p(D|\theta)\, p(\theta)$$

- This is known as **Maximum a Posteriori (MAP)** estimation [4]

[4] You are going to learn more Posterior distributions next year in Machine Learning!

Maximum a Posterior - Example

Example

Given a coin, you were assigned the task of figuring out whether the coin will land on its head or tails. You were asked to build a probabilistic model (i.e. with confidence)

- Suppose we want to utilise our prior belief that coins are typically fair
- $p(\theta)$ would peak around $\theta = 0.5$
- Let's use

$$p(\theta) = b\,\theta\,(1-\theta)$$

where b is a normalising factor so the area under the curve is equal to 1

Maximum a Posterior - Example

- **Likelihood:**

$$p(D|\theta) = \binom{N}{D} \theta^{D} (1-\theta)^{N-D}$$

- **Prior:**

$$p(\theta) = b\,\theta\,(1-\theta)$$

- **Posterior:**

$$p(\theta|D) = p(D|\theta)\,p(\theta) = \binom{N}{D} \theta^{D} (1-\theta)^{N-D}\, b\,\theta\,(1-\theta)$$

Maximum a Posterior - Example

- Take the natural logarithm and derivate [same recipe as in MLE]

$$p(D|\theta)\,p(\theta) = \binom{N}{D}\theta^D(1-\theta)^{N-D}\,b\,\theta\,(1-\theta)$$

$$\ln p(D|\theta)\,p(\theta) = \ln\binom{N}{D} + D\ln\theta + (N-D)\ln(1-\theta) + \ln b + \ln\theta + \ln(1-\theta)$$

$$\frac{d}{d\theta}\ln p(D|\theta)\,p(\theta) = D\frac{1}{\theta} - (N-D)\frac{1}{1-\theta} + \frac{1}{\theta} - \frac{1}{(1-\theta)}$$

Maximum a Posterior - Example

- Set the derivative to 0 and solve for θ_{MAP}

$$D\frac{1}{\theta_{MAP}} - (N-D)\frac{1}{1-\theta_{MAP}} + \frac{1}{\theta_{MAP}} - \frac{1}{(1-\theta_{MAP})} = 0$$

$$\frac{D+1}{\theta_{MAP}} - (N-D+1)\frac{1}{1-\theta_{MAP}} = 0$$

$$\frac{(D+1)(1-\theta_{MAP}) - (N-D+1)\theta_{MAP}}{\theta_{MAP}(1-\theta_{MAP})} = 0$$

$$\theta_{MAP} = \frac{D+1}{N+2}$$

- The prior added two 'virtual' coin tosses, one with heads and one with tails. Note that for $D = N = 0$, it defaults to our prior knowledge, i.e. $\theta_{MAP} = \frac{1}{2}$.

Conclusion

- Probabilistic models encode randomness in the data
- They provide model uncertainty
- Parameters of the model are tuned using estimators
- **Maximum Likelihood Estimation (MLE)** is a recipe used for training model parameters
- MLE does not encode our prior knowledge of possible parameters
- **Maximum a Posteriori (MAP)** maximises likelihood along with prior

Further Reading

- Probability and Statistics for Engineers and Scientists
  Walpole et al (2007)
  - Section 3.1
  - Section 3.2
  - Section 4.1
  - Section 4.2
- Statistical Learning Methods
  Russell and Norvig (2003)
  - Chapter 20 (p. 712 - 720)