COMS21202: Symbols, Patterns and Signals
Probabilistic Data Models

Dima Damen
Dima.Damen@bristol.ac.uk

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

February 9, 2019

Data Modelling

- Deterministic models do not explicitly model uncertainties or 'randomness' in data
- Variability of inferences derived from the data is not included
- In many tasks, we benefit from modelling uncertainty and randomness
- This is explicit in **Probabilistic Models**

Back to Fish - Discrete

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

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

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, usually close to zero

Back to Fish - Continuous

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

- To model the random variable, we measure the difference between the *predicted* and *measured* weight values
- Modelled using a probability distribution for ϵ,
  - by a uniform distribution
  - by a normal distribution
  - ...
- In the next slides, we will make the *logical* simplification (weight = 0 when length = 0)
- As a conclusion, the y-intercept can be set to zero, and

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

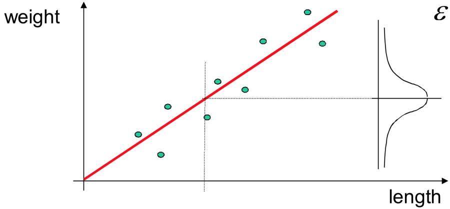
Back to Fish - Continuous

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

This is a model with one parameter, apart from the uncertainty

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

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



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
- **MLE** is an extremum estimator obtained by maximising an objective function of $\boldsymbol{\theta}$

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

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

- **Note:** this is different than maximising the function (i.e. finding the maximum value [$\mathit{max}\, f(\theta)$])
- Tuning the parameter is then equal to finding the maximum argument *arg max*

Dima Damen
Dima.Damen@bristol.ac.uk

Maximum Likelihood Estimation

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 parameter a

    $$p(D|a)$$

  - using observed data, find paramter value which maximises the conditional probability (i.e. the likelihood)

    $$a_{ML} = arg\, max_a p(D|a)$$

Maximum Likelihood Estimation

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 :

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

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

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

For a large sample:

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

Maximum Likelihood Estimation

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

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

Maximum Likelihood Estimation

To tune the parameter, i.e. find the ML parameter,

$$
\begin{aligned}
a_{ML} &= arg\ max_a\, p(D|a) \\
&= arg\ max_a \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(y_i - ax_i)^2}{\sigma^2}} \\
&= arg\ max_a \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(y_i - ax_i)^2}{\sigma^2}}\right) \\
&= arg\ max_a \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(y_i - ax_i)^2}{\sigma^2}}\right) \\
&= arg\ max_a \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\sigma}} - \frac{(y_i - ax_i)^2}{2\sigma^2} \\
&= arg\ max_a \sum_{i=1}^{N} -(y_i - ax_i)^2 \qquad \text{(remove constants)} \\
&= arg\ min_a \sum_{i=1}^{N} (y_i - ax_i)^2
\end{aligned}
$$

Data Modelling - Deterministic vs Probabilistic

- Deterministic Least Squares:

$$a_{LS} = arg\,min_a\, R(a) = arg\,min_a \sum_i (y_i - a\,x_i)^2$$

- Probabilistic Maximum Likelihood:

$$a_{ML} = arg\,min_a \sum_i (y_i - a\,x_i)^2$$

- same answer, different view
- **Note:** ML answer here assumes uncertainty is normally distributed

Data Modelling - Deterministic vs Probabilistic

In both cases,

$$a_{ML} = arg\, min_a \sum_i (y_i - a\, x_i)^2$$

To find the minimum, find the derivative

$$\frac{d}{da} \sum_i (y_i - ax_i)^2 = -2 \sum_i x_i (y_i - ax_i)$$

and equate it to zero

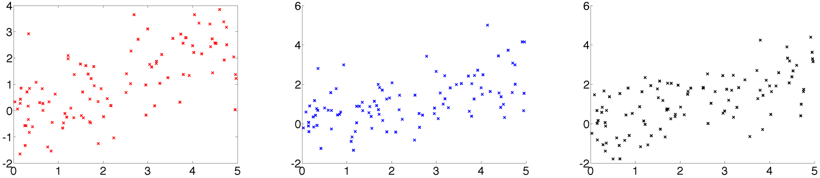
$$-2 \sum_i x_i (y_i - a_{ML} x_i) = 0$$

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

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

Data Modelling - Deterministic vs Probabilistic

- so why to take the probabilistic approach?
- **Probabilistic Models** can tell us more
- For example: how much does a_{ML} vary if it is computed for many data samples? How reliable is it?

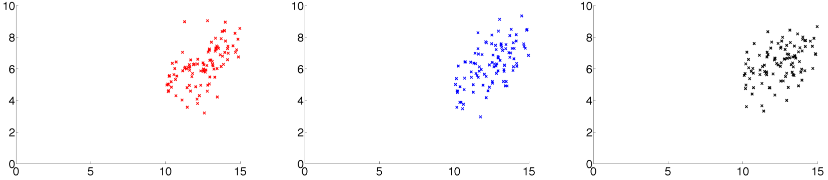


$$a_{ML} = (0.51, 0.49, 0.52)$$

Dima Damen
Dima.Damen@bristol.ac.uk

Data Modelling - Deterministic vs Probabilistic

- so why to take the probabilistic approach?
- **Probabilistic Models** can tell us more
- For example: how much does a_{ML} vary if it is computed for many data samples? How reliable is it?



$$a_{ML} = (0.50, 0.50, 0.51)$$

Dima Damen
Dima.Damen@bristol.ac.uk

Data Modelling - Deterministic vs Probabilistic

- so why to take the probabilistic approach?
- **Probabilistic Models** can tell us more
- For example: how much does a_{ML} vary if it is computed for many data samples? How reliable is it?
- For M different samples

$$Var(a_{ML}) = \frac{1}{M-1}\sum_{j=1}^{M}(a_{MLj} - \overline{a_{ML}})$$

- If

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

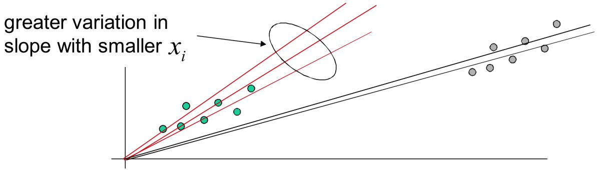
- Then for the same values x_i

$$Var(a_{ML}) = \frac{\sigma^2}{\sum_i x_i^2}$$

Data Modelling - Deterministic vs Probabilistic

$$Var(a_{ML}) = \frac{\sigma^2}{\sum_i x_i^2}$$

- Variance is thus dependent on input variables



Maximum Likelihood Estimation - General

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

$$
\begin{aligned}
\theta_{MLE} &= arg\ max_{\theta}\ p(D|\theta) \\
&= arg\ max_{\theta}\ \ln p(D|\theta) \\
&= arg\ min_{\theta}\ -\ln p(D|\theta)
\end{aligned}
$$

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

Maximum Likelihood Estimation - Ex1

MLE Recipe - Ex1

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

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 of heads to the sample

Probabilistic Model - Ex2 - again

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)

What if you chose another model?

- **Data:** head/tail binary attempts (of size N)
- **Model:** Normal distribution
- **Model Parameters:** mean μ - assume σ is a constant

Probabilistic Model - Ex2 - again

Assume $D = \{d_1, d_2, \cdots d_N\}$ are *noisy* measurements of an actual signal $\theta = \mu$, where noise is Gaussian,

$$p(D|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d_i-\theta)^2}{2\sigma^2}}$$

i.e. $D = \{0, 0, 1, 1, 1, \cdots\}$ where 0 represents tails and 1 represents heads...

Probabilistic Model - Ex2 - again

- Take the natural logarithm and derivate

$$p(D|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(d_i-\theta)^2}{2\sigma^2}}$$

$$\ln p(D|\theta) = N \ln \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^{N} -\frac{(d_i-\theta)^2}{2\sigma^2}$$

$$\frac{d}{d\theta} \ln p(D|\theta) = \sum_{i=1}^{N} -\frac{2(d_i-\theta)(-1)}{2\sigma^2}$$

Probabilistic Model - Ex2 - again

- Set the derivative to 0 and solve for θ

$$\sum_{i=1}^{N} \frac{(d_i - \theta_{ML})}{\sigma^2} = 0$$

$$\sum_{i=1}^{N} d_i - N\theta_{ML} = 0$$

$$\theta_{ML} = \frac{1}{N} \sum_{i=1}^{N} d_i$$

$$\theta_{ML} = \overline{d}$$

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)

- Use binomial distribution for likelihood

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

  where D is the number of success (i.e. heads)
- Use Gaussian distribution for likelihood

$$\theta_{ML} = \frac{1}{N} \sum_{i=1}^{N} d_i$$

  where $d_i = 1$ if success (i.e. heads) or $d_i = 0$ if failure (i.e. tails)
- same answer, different view

Probabilistic Model - Likelihood and Prior

- MLE ignores any prior knowledge we may have about θ
- If we have prior knowledge about values that θ is likely to have, then we can built this into MLE

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

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

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:**

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

Maximum a Posterior - Example

- Take the natural logarithm and derivate

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

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

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

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

Conclusion

- Probabilistic models encode randomness in the data
- They enable predicting confidence (as a probability)
- Parameters of the model are tuned
- **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)