FIT5201 - Data analysis algorithms

Online quiz 1 has been released

Deadline: 19:59:59 14th , August (8pm next Wednesday)

FIT5201 - Data analysis algorithms

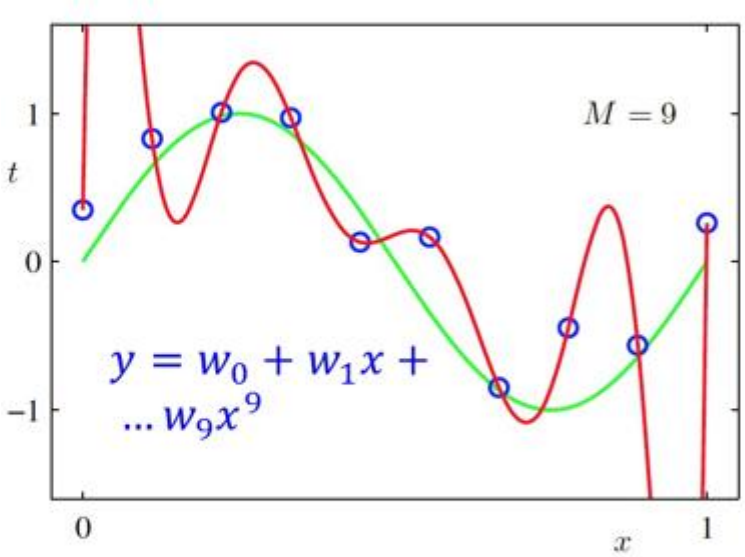
Module 1: Elements of Machine Learning

Part B:

- **Probabilistic Machine Learning**
  - □ Learn basics of **probability theory**
  - □ Learn how probability theory can help to capture the key concept, "**uncertainty**", in machine learning
    - Frequentist (bootstrap)
    - Bayesian (probability)
  - □ Learn **Bayesian approach** to **machine learning**

Uncertainty in Machine Learning (ML)

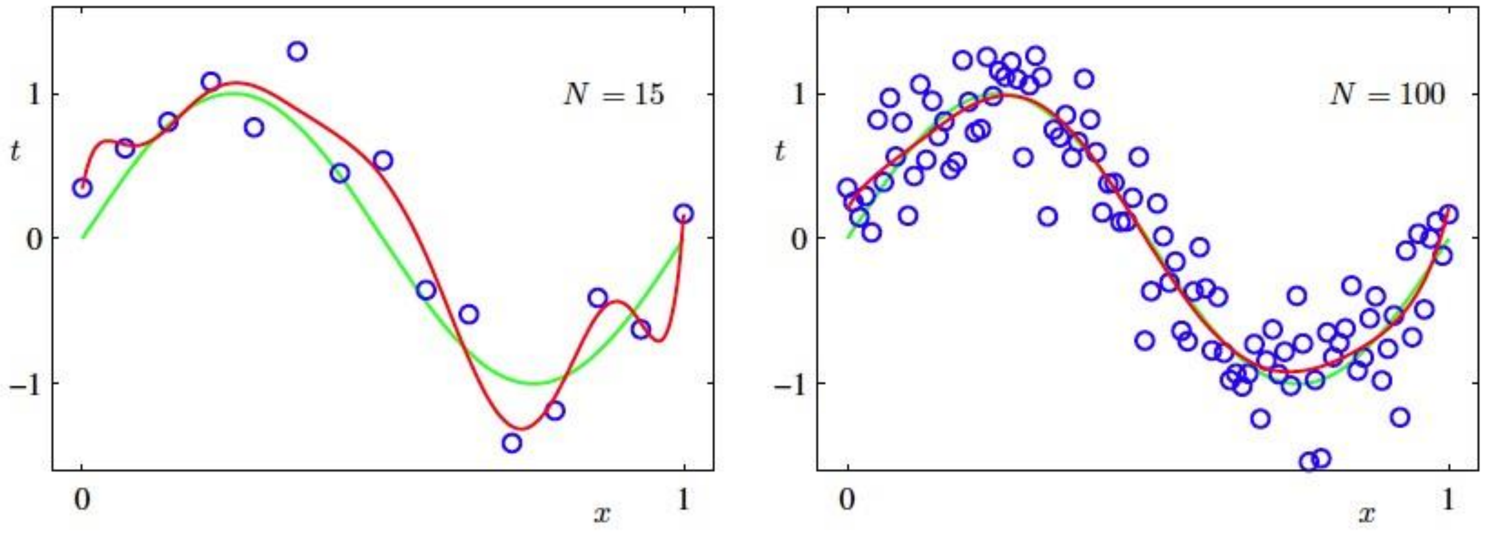
- ❑ A key, common concept in ML.
- ❑ What is the uncertainty in machine learning, exactly?



- ❑ The reasons?

Uncertainty in Machine Learning (ML)

- ❑ Two key factors
  - o Noise
  - o Finite size of data sets



Probability Theory

- ❑ Why study probability theory?
  - □ It provides a framework for measuring and manipulating ”uncertainty”.
  - □ What we are going to do?
    - Review the basics of probability theory
    - Use probability theory to capture our uncertainty when learning models and making predictions
      - ➢ Frequentist
      - ➢ Bayesian statistics

Basics of Probability Theory

Coin tossing example

How to model and represent the coin toss event mathematically?

Central subjects in probability theory include discrete and continuous random variables, probability distributions, and stochastic processes, which provide mathematical abstractions of non-deterministic or uncertain processes

Basics of Probability Theory

Coin tossing example

Random Variable: X

1. Create a **random variable** X to denote the outcome of tossing a coin: H(Head), T(Tail)

2. Domain of X: $D_X := \{H, T\}$.

3. In the limit that the total number of trials goes to infinity, we can define the probability of an event to be a fraction: (times that event occurs / total number of trials) uncertainty caused by finite dataset!

4. The probability distribution, p(.) for X is a function which maps each value of X to its probability.

   p(X=H)=0.2

Basics of Probability Theory

Coin tossing example

Random Variable: X

❑ Basic rules of probabilities:

➢ (1) Probabilities must lie in [0,1]; and

➢ (2) The sum of the probabilities of each possible individual outcome must be 1 (e.g. $p(X = H) + p(X = T) = 1$): $\sum_{a \in \mathrm{Dx}} p(X = a) = 1$

❑ Simply, we use $\sum_x p(x) = 1$, where the lowercase letter denotes the values.

Condition: there is no confusion about the domain of the random variable

Basics of Probability Theory

Coin tossing example

Random Variables: X, Y, Z

1. Three random variables
   X: tossing the coin for the first time ($D_X = \{H, T\}$)
   Y: tossing the coin for second time ($D_Y = \{H, T\}$)
   Z: tossing the coin for two times ($D_Z = \{HH, HT, TH, TT\}$)

2. Conditional probability
   The probability of event Z occurring given that event X occurs

$$p(Z = HH | X = H)$$

$$p(Y = H | X = H) = p(Y = H)$$

Basics of Probability Theory

Coin tossing example

Random Variables:
X, Y, Z

3. Joint probability

Given two random variables X and Y, the joint probability distribution p(X,Y) gives the probability of possible combination of values for these two random variables

$$\sum_x \sum_y p(x,y) = 1$$

$$p(Z = HH, X = H)$$
$$= p(Z = HH | X = H) p(X = H)$$

Independence

$$p(X = H, Y = H)$$
$$= p(X = H) p(Y = H)$$

Basics of Probability Theory

❑ Sum and Product Rules

- □ Sum Rule:

  $p(x) = \sum_y p(x, y)$, $p(x)$ is the marginal probability and simply read as " the probability of x "

- □ Product Rule:

  $p(x, y) = p(y|x)p(x)$, $p(x, y)$ is the joint probability and simply read as " the probability of x and y "

❑ Bayes Theorem

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$where\ p(x) = \sum_y p(x|y)p(y),$ which can be seen as a normalization constant required to ensure that the sum of $p(y|x)$ over all values of Y equals one

Basics of Probability Theory

❑ Sum and Product Rules

- □ Sum Rule:

  $p(x) = \sum_y p(x, y), p(x)$ is the marginal probability and simply read as “ the probability of x ”

- □ Product Rule:

  $p(x, y) = p(y|x)p(x), p(x, y)$ is the joint probability and simply read as “ the probability of x and y ”

❑ Bayes Theorem

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

where $p(x) = \sum_y p(x|y)p(y)$, *which can be seen as a normalization constant required To ensure that the sum of* $p(y|x)$ *over all values of Y equals one*

$$\begin{aligned} p(y|x) &= \frac{p(x,y)}{p(x)} \\ &= \frac{p(x|y)p(y)}{p(x)} \\ &= \frac{p(x|y)p(y)}{\sum_y p(x,y)} \\ &= \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)} \end{aligned}$$

Measuring Uncertainty using Probability Theory: Learning about a coin

Learning about a coin

- **D = {H, T, T, T, T, H, T, T, H, T}**
  - D contains the outcomes of 10 independent tosses of the same coin.
  - The coin might be damaged
  - We are asked to build a model for this coin tossing to predict the probability that next coin toss will be a head.

Learning about a coin

- D = {H, T, T, T, T, H, T, T, H, T}

Every toss is random. So there is some probability to be observed. Let us denote that the given coin has probability w of coming up a head.

For the i-th coin toss, we write $p(H) = w$, where w is the parameter in [0,1]. Since the sum of probabilities of all outcomes must be 1, $p(T) = 1 - w$.

However, we DON'T know w. To estimate the information about w, we need to use the observed sequence D.

Maximum Likelihood

- ❑ **D = {H, T, T, T, T, H, T, T, H, T}**

To denote that the probability distribution of D depends on w, we write

$$p(D|w)$$

What's the probability that we observe D? That depends on w.

Maximum Likelihood

❑ **D = {H, T, T, T, T, H, T, T, H, T}**

Find w's such that $p(D|w)$ is as high as possible (i.e. Maximise the likelihood of our observation)

How to calculate this probability? $(x_i = H \text{ or } T)$

$$p(D|w) = \prod_{i=1}^{10} p(x_i|w)$$

Notice $p(x_i|w)$ is w, if the i-th toss is "H", and $(1 - w)$ if it is "T".
Whatever w is, the probabilty of getting D can be expressed as:

$$\prod_{i=1}^{10} p(x_i|w) =$$

$$w(1-w)(1-w)(1-w)(1-w)w(1-w)(1-w)w(1-w) = w^3(1-w)^7$$

Maximum Likelihood

- ❑ **D = {H, T, T, T, T, H, T, T, H, T}**

$$p(D|w) = w^3(1-w)^7$$

If w had been either 0 (heads impossible) or 1 (tails impossible), then the probability of getting D would have been 0.

If w had been ½ (a fair coin), then the probability of getting D would have been 1/1024.

Maximum Likelihood

❑ D = {H, T, T, T, T, H, T, T, H, T}

How can we find w that maximise the probability ?

$$p(D|w) = w^3(1-w)^7$$

Usually easier to apply a monotone transformation to the likelihood that converts multiplication to addition - the logarithm function:

Find w that maximise the following

$$\ln(p(D|w)) = 3\ln(w) + 7\ln(1-w)$$

Maximum Likelihood

- ❑ Likelihood function
  - ➢ $p(D|model)$, usually maximize its logarithm
  - ➢ error function: negative log of the likelihood function

Maximum Likelihood

❑ **D = {H, T, T, T, T, H, T, T, H, T}**

From

$$\ln(p(D|w)) = 3\ln(w) + 7\ln(1-w)$$

How can we find such w that maximises the above function?

It can be solved by taking the derivative setting the resulting expression equal to 0: that is,

$$\frac{\partial \ln(p(D|w))}{\partial w} = \frac{3}{w} - \frac{7}{1-w} = 0$$

$$\rightarrow w = 0.3$$

What is the probability that next coin toss will be a **head**? **0.3**

Maximum Likelihood

❑ **D = {H, T, T, T, T, H, T, T, H, T}**

From

$$\ln(p(D|w)) = 3\ln(w) + 7\ln(1-w)$$

Naive?

How can we find such w th ction?

It can be solved by taking ulting expression equal to 0: that is,

$$\frac{\partial \ln(p(D|}{\partial w} - \frac{}{w} \quad \frac{}{1-w} = 0$$

$$\rightarrow w = 0.3$$

What is the probability that next coin toss will be a **head**? **0.3**

Learning about a coin

❑ Summary

- The prediction problem (probability that next coin is H or T) is formulated as finding an optimal coin toss parameter w that maxmises the probability that we observe D

- To find such w, we can use logarithm and derivative calculation

- Once we obtain such w, we can use it for prediction purposes.

- The parameter w clearly depends on the dataset D. So w can be changing if we have a different dataset D'.

Bootstrap

Bootstrap for Quantifying Uncertainty

- ❑ Imagine
  - o We only have D, and our goal is to fit a model with parameter w to the given dataset.
  - o We found w that maximises the probability of observation D.
  - o We wonder if w would change if we have an alternate dataset D'
  - o If we do this exercise for several alternate datasets, then we will a *distribution* over estimates for w .
  - o If this distribution is higher, the more uncertainty on w.

Now, the problem is how to get the alternate datasets?

Bootstrap for Quantifying Uncertainty

Now, the problem is how to get the alternate datasets?

clue: generate the alternate datasets from the original dataset. But how?

Solution: resampling with bootstrap strategy

Basic idea of bootstrap: approximate the real distribution in the infinity situation with the empirical distribution

Bootstrap Example

❑A bootstrap sample is a random sample conducted with replacement

1. Randomly select an observation from the original data
2. Write it down
3. Put it back (i.e. Any observation can be selected more than once)

Repeat these steps 1-3 N times; N is the number of observations in the original sample

Final Result: One "bootstrap sample" with N observations

0	**48** **37** **24**	**3** **1** **5**

Original Data

N

Target	Age	Inquires
0	48	3
0	37	1
0	37	1
	…	
0	24	5

Summary

Classical or frequentist interpretation of uncertainty

1. View probabilities in terms of the frequencies of random, repeatable events, in the limit of infinity.

2. Use bootstrap sample strategy to simulate infinity based on the available data.

3. Learn the model by maximizing the likelihood of every single bootstrap data set.

4. Evaluate the uncertainty by looking at the variability of predictions between the different bootstrap data sets.

Bayesian Approach to Machine Learning

(use probabilities to quantify the uncertainty directly)

What is Bayesian Approach to ML

❑ Significance of Bayes' Theorem

- Bayes Theorem is the fundamental result of probabiliy theory
- Bayes Theorem is the basis of a branch of ML

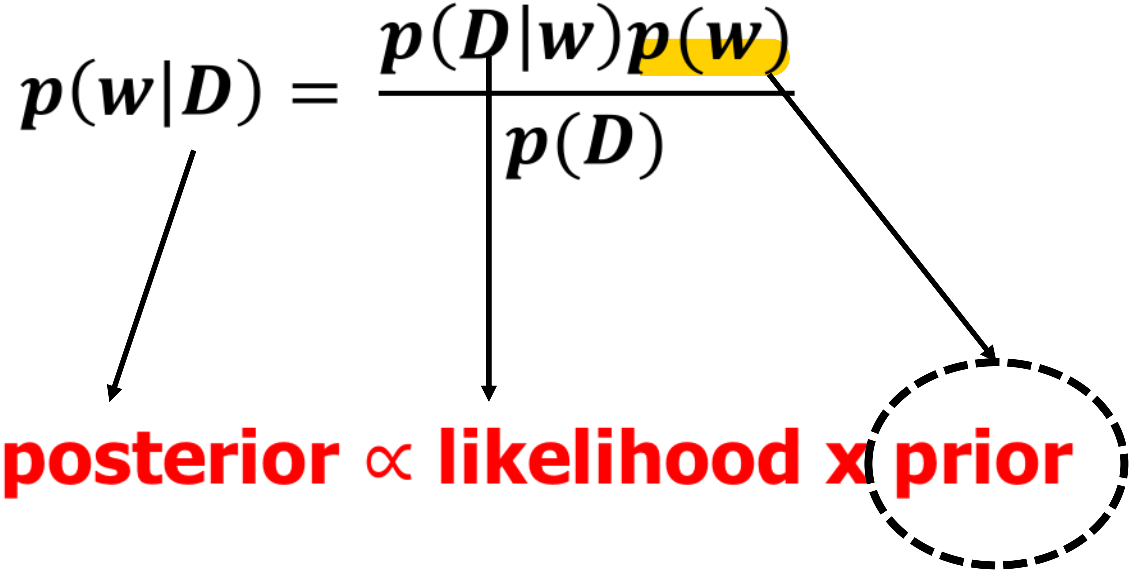
❑ Bayes' Theorem

$$\boldsymbol{p(M|D) = \frac{p(D|M)p(M)}{p(D)}}$$

- $p(M|D)$: the probability of the model M given the data D
- $p(D|M)$: the probability of data D given by the model M
- $p(M)$: the prior probability of the model M
- $p(D)$: the probabilty of the data D

What is Bayesian Approach to ML

- Given the definition of likehood, we can state Bayes' theorem in words:



$$p(\boldsymbol{w}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{D})}$$

posterior ∝ likelihood x prior

What is Bayesian Approach to ML

For a model w

1. Initialize $p(w)$ with people's prior knowledge (subjective)

2. Once you observe a new data D, compute the likelihood $p(D|w)$

3. Compute the posterior $p(w|D)$ according to Bayesian Theorem

4. Update the prior knowledge $p(w) = p(w|D)$

Do step 2 to 4 iteratively

Encoding Prior Knowledge

- What's the limitation of the above model?

  - Imagine that **D = {T, T}** with the same coin! Then, what's the probability that the next coin toss will be a head?

$$\mathbf{w}_{\text{MLE}} = \frac{|\boldsymbol{H}|}{|\boldsymbol{H}| + |\boldsymbol{T}|} = \frac{\mathbf{0}}{\mathbf{2} + \mathbf{0}} = \mathbf{0}$$

  - The result $w = 0$ is unreasonable.

  - Fortunately, from your experience, most coins are about 50-50.

  - Use **Bayesian modelling** to encode this **prior experience or knowledge about the parameter: $p(w)$**

Encoding Prior Knowledge

- ❑ Encoding prior knowledge about a model (or parameter):
  - $p(w)$: the **prior probability** of parameter w

- ❑ But we don’t know what specific value for w that we should use. So we will integrate over all possible values of w:

$$\int_X p(w = x)dx = 1, w \in [0,1]$$

  - o $p(w = x)$: the probability that x is w but without any observation (i.e. prior knowledge about w at x)

- ❑ Simple,

$$\int p(w)dw = 1$$

Encoding Prior Knowledge

The probablity distributions $p(w)$ follow Beta distribution,

$$\mathbf{Beta}(\boldsymbol{w}|\boldsymbol{a},\boldsymbol{b}) \propto \boldsymbol{w}^{\boldsymbol{a}-\mathbf{1}}(\mathbf{1}-\boldsymbol{w})^{\boldsymbol{b}-\mathbf{1}}, \boldsymbol{w} \ \in [\mathbf{0},\mathbf{1}]$$

We have seen this before: $\boldsymbol{w}^{\boldsymbol{a}-\mathbf{1}}(\mathbf{1}-\boldsymbol{w})^{\boldsymbol{b}-\mathbf{1}}$ (very similar to the estimate in MLE). Remember: $\boldsymbol{w}^{\mathbf{3}}(\mathbf{1}-\boldsymbol{w})^{\mathbf{7}}$

In the MLE estimate, we had a similar function form but (a-1) and (b-1) were substituted by the number of Heads and Tails respectively!

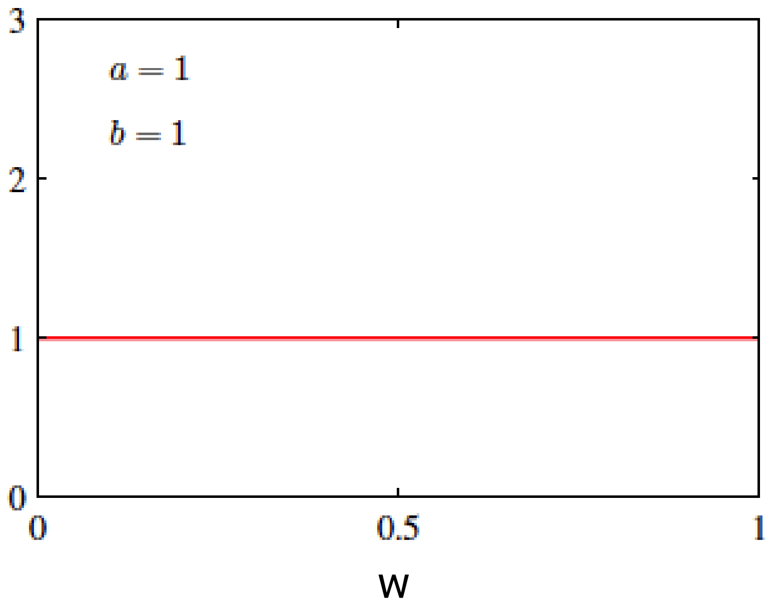
Highlight that Beta distribution encodes prior knowledge about parameter w **before** actually observing data.

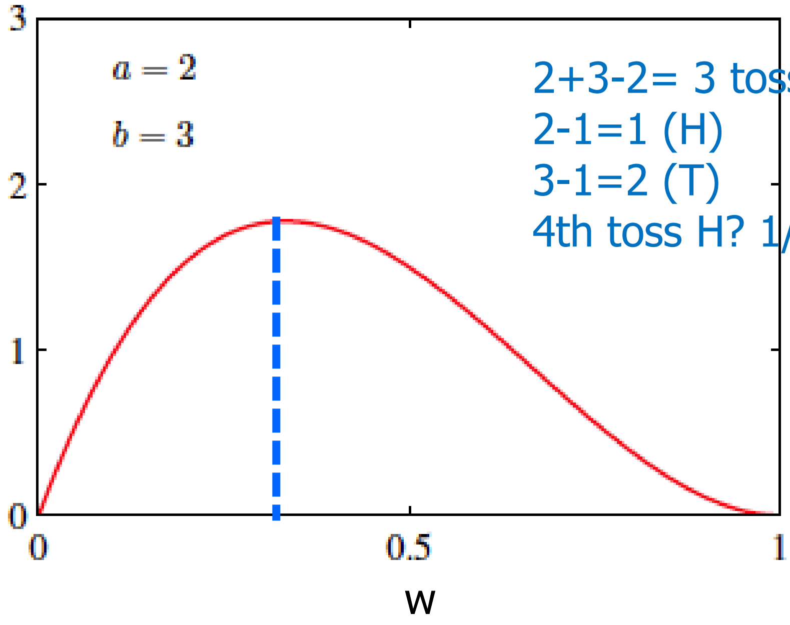
Encoding Prior Knowledge

Prior knowledge (or belief),

$$\mathbf{p(w)} \propto \mathbf{Beta}(\boldsymbol{w}|\boldsymbol{a},\boldsymbol{b}) \propto \boldsymbol{w}^{\boldsymbol{a}-1}(\mathbf{1}-\boldsymbol{w})^{\boldsymbol{b}-1}, \boldsymbol{w} \in [\mathbf{0},\mathbf{1}]$$

How to interprete: (a-1) and (b-1) are the number of Heads and Tails I think we would see, if we made (a+b-2) coin tosses!





2+3-2= 3 tosses
2-1=1 (H)
3-1=2 (T)
4th toss H? 1/3

What is Bayesian Approach to ML

For any x, I want to estimate $p(w = x|D)$, where D is a randome variable that models the observed data: simply we denote this $p(w|D)$.

Bayes' Theorem:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

- $p(D|w)$: We know this one! This is used to estimate parameter w in MLE
- $p(w)$: We know this one too! This is prior belief – probability distribution over parameter w
- $p(D)$: that's the probability of the observed data D

What is Bayesian Approach to ML

Bayes' Theorem:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

Let's put together the numerator!

$$p(D|w) = w^{|H|}(1 - w)^{|T|} \text{(by MLE)}$$
$$p(w) \propto w^{a-1}(1 - w)^{b-1} \text{(by prior belief)}$$

So we get the numerator:

$$\begin{aligned} p(w|D) &\propto w^{|H|+a-1}(1 - w)^{|T|+b-1} \\ &\propto \mathrm{Beta}(w||H| + a, |T| + b) \end{aligned}$$

What is Bayesian Approach to ML

Bayes' Theorem:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

$\boldsymbol{p(D)},$ that's the probability of the observed data.

$$p(D) = \int p(D,w)dw = \int p(D|w)p(w)dw$$

The product rule

Sum Rule for discrete values for y: $p(x) = \sum_y p(x,y)$

Sum Rule for discrete continues values for y: $p(x) = \int_y p(x,y)dy$

What is Bayesian Approach to ML

Prediction problem:

How can we predict the probability that next coin toss is Head, given observation D using Bayesian approach:

$$\boldsymbol{p(H|D)}$$

A toss depends on w, but we don't know what specific value for w we should use. So integrate over all possible value of w.

$$p(H|D) = \int_0^1 p(H, w|D)dw$$

What is Bayesian Approach to ML

$$p(H|D) =$$

$$\int_0^1 p(H, w|D)dw = \int_0^1 p(H|w)p(w|D)dw$$

How? Product Rule + Conditional Independence Assumption!

If we know w, then the result for H only dependes on that knowledge, all other information is no longer important, i.e.:

$$p(H|w, D) = p(H|w)$$

What is Bayesian Approach to ML

Continuing by plugging in all formula we have …

$$p(H|D) = \int_0^1 p(H, w|D)dw = \int_0^1 p(H|w)p(w|D)dw =$$

$$\int_0^1 w^{|H|}(1-w)^{1-|H|}\ p(w|D)dw =$$

$$w\mathrm{Beta}(w||H| + a, |T| + b) =$$

$$E[w|D] = \frac{|H| +\ a}{|H| + |T| +\ a +\ b}$$

$$\left(E[w] = \ \frac{a}{a+ b}\ \text{from Beta}(w|a, b)\right)$$

Summary of Bayesian Approach

This shows what?

We can predict the class of a new sample (e.g. what will be the result of a new coin toss) based on the prior and the data.

$$p(H|D) = \frac{|H| + a}{|H| + |T| + a + b}$$

- a = b = 100: A strong prior belief that the coin is fair
- a = b = 2: A weak prior belief that the coin is fair
- a=100, b=1: For a coin biased towards heads

- If D = {T, T} , probability that next coin is H?

Wrap up the lecture

- ❑ Learned basics of probability theory
  - o Probabilities must lie in [0,1]
  - o The sum of probabilities of a random variable must be 1
  - o Sum Rule: $\boldsymbol{p(x) = \sum_y p(x,y)}$
  - o Product Rule: $\boldsymbol{p(x,y) = p(y|x)p(x)}$
  - o Conditional probability: $\boldsymbol{p(y|x)}$
  - o Joint probability: $\boldsymbol{p(x,y) = p(y|x)p(x)}$
  - o Bayes' theorem: $\boldsymbol{p(y|x) = \frac{p(x|y)p(y)}{p(x)}}$

Wrap up the lecture

❑ Parameter estimation

□**Maximum Likelihood Estimation (MLE)** is a central parameter estimation is machine learning

□MLE principle says to choose the maximum likelihood of parameter(s) that maximise the probability of the observed data

□That is, **MLE** can be used to determine best model parameter(s) that fit the given data.

Wrap up the lecture

- ❑ Learned how probability theory can help to capture the uncertainty in machine learning through the coin tossing problem: $p(w|D)$ - evaluate uncertainty in w after we have observed D

- ❑ Learned how Bayesian approach to machine learning can maximise the posterior distribution $p(w|D)$ based on the prior knowledge and likelihood.

- ❑ Learned how a Bayesian approach can be used for a prediction problem.
  - o Never forget how it work!!!!

Tutorial (Week 2)

- ❑ Practice Bootstrap Techniques
  - □ We will implement the Bootstrap technique to assess variations in the prediction of KNN classifier.

What will we learn in Week 3

- ❑ **Part A in Module 2**
  - □ Linear Basis Functions
  - □ Optimising Error Functions
  - □ Regularisation algorithms

THANK YOU
for your attention!