Week 2 - Probabilistic Machine Learning

Different prospectives of uncertainty: Frequentist v.s Bayesian

- **Frequentist**: define a notion of probability by doing numerous independent trial
    - Make use of Central Limit Theorem (CLT)
    - No event can be repeated numberous times (earthquake)
    - Can not reasonably explain rare event without sufficient data
- **Bayesian**: quantify a belief(uncertainty) of an event into probability and then make precise revision of it in the light of new evidence, as well as to be able to take optimal actions or decisions as a consequence. (Knowing the probability of consequence before taking the action)
    - Quantify uncertainty by the previous belief (prior) and the data in hand before getting sufficient data
    - Easily to apply but difficult to interpret (belief based on mathematical convenience)
    - The prior initialise with prior knowledge (subjective)

> Bayesian approach in machine learning use the machinery of probability theory to describe the uncertainty in model parameters, or indeed in the choice of model itself.

---

1. Probability Theory

A **framework** for measuring and manipulating uncertainty.

- Central subjects in probability theory include **discrete** and **continuous random variables**, **probability distribution**, and **stochastic processes**, which provide mathematical abstractions of nondeterministic or uncertain processes
- **Basic rules**:
    - Probabilities must lie in $[0, 1]$
    - The sum of the probabilities of each possible individual outcome must be 1
- **Conditional probability**:
    - Dependent: $p(A|B) = \dfrac{p(A, B)}{p(B)}$
    - Independent: $p(A|B) = p(A)$
- **Joint probability**:
    - Dependent: $p(A, B) = p(A|B) \times p(B)$
    - Independent: $p(A, B) = p(A) \times p(B)$
    - $\sum_A \sum_B p(A, B) = 1$
- **Marginal probability**: $\sum_B p(A|B) = p(A)$

2. Bayes' theorem

It converts a **prior** probability into a **posterior** probability by incorporating the **evidence** provided by **the observed data**.

Making inferences about the quantities of the parameter $\boldsymbol{w}$:

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

- **Prior** $p(\boldsymbol{w})$: Assumption about the event before observing the data (D)

  > The prior will be updated by each assessment of posterior probability.

- **Likelihood** $p(D|\boldsymbol{w})$: The observed data (**evidence**) is expressed through the conditional probability

  > Given the $\boldsymbol{w}$, the probability of the occurrence of $D = \{x_i, i \in N\}$. Then,
  >
  > By using convolution equation: $p(D|\boldsymbol{w}) = \prod_{n=1}^{N} p(x_i|\boldsymbol{w})$
  >
  > We need to find the best $\boldsymbol{w}$ to maximise $p(D|\boldsymbol{w})$.

- **Posterior** $p(\boldsymbol{w}|D)$: Evaluate the uncertainty in $\boldsymbol{w}$ after we have observed D

  > Given the observed data D, the probability of $\boldsymbol{w}$. It in turn updates our previous assumption (prior) and move on to the next assessment.

- **Normalization** constant $p(D)$: $\sum_{\boldsymbol{W}} p(D|\boldsymbol{w})p(\boldsymbol{w})$ ($\boldsymbol{W}$ includes all possible $\boldsymbol{w}$)

  > It can be treated as a constant.

3. Maximum likelihood Approach (MLE)

- It is a widely used **frequentist estimator**.
- **Error function**: To find the best $\boldsymbol{w}$ to maximise $p(D|\boldsymbol{w})$, we need to use **the negative log of the likelihood function** which is call **Error function**, also called **Objective function**.

> **Why use the negative logarithm**($-\log x$)?
>
> The negative logarithm is a monotonically decreasing function, maximising the likelihood is equivalent to minimising the error. It shift the problem from maximising the likelihood to minimise the error function, which is more easy to solve.

- **Example**: let say training a model for flipping coin event which we assume it only involve single parameter a, let H as head.
  - let $p(x = H) = a$.
  - Assume now we have D data set with only 10 data, 3 times show head, 7 times show tail. we can have the equation by applying Bayes' theorem: $p(D|a) = a^3(1-a)^7$.
  - Then we want to use maximum likelihood to find the best guess of a, we use the negative logarithm: $\ln\{p(D|a)\} = 3\ln(a) + 7\ln(1-a)$.
  - With this equation, we differentiate it with respect to the parameter a:
    $$\frac{\partial \ln\{p(D|a)\}}{\partial a} = \frac{3}{a} - \frac{7}{1-a}$$

  - Find the stationary point by setting it equaled to zero: $\frac{\partial \ln \{p(D|a)\}}{\partial a} = 0 \Rightarrow a = 0.3.$

- **The limitation of maximum likelihood approach**: Without **sufficient training data**, maximum likelihood **systematically underestimates the variance of the distribution**. The ideal size of the training data for this approach should be **infinity**. However, it is **impractical** to have such volume of data for training model.

- **Adding bias**: Also, we can find that the outcome of this likelihood can contribute to the posterior $p(a|D)$ which feed to the prior $p(a')$ in next assessment of uncertainty. **It involves bias into the model!**

Bootstrap:

- If we only have limit volume data $X = \{x_1, x_2, \cdots, x_N\}$ (N data points), but we have to find out the accuracy of our parameter estimates. Then, we can use Bootstrap to create new data set by drawing N data points (the same size as the source) at radom from X , **with replacement**.
- It make use of the CLT: the more subset you sample, the average of all subset mean are more close to the ensemble mean.
- This process can be repeated multiple times to estimate **the variability of the prediction** between the **different bootstrap data sets** at **different parameter estimates**.
- **The higher the variability, the higher the given uncertainty of the parameter estimates**.
- It **simulate infinity** based on the available data.

4. Bayesian Approach

- Step by step:
  - Encode prior $p(\boldsymbol{w})$ with previous knowledge
  - Once observe a new data D, compute the likelihood $p(D|\boldsymbol{w})$
  - Compute the posterior $p(\boldsymbol{w}|D)$ according to Bayesian Theorem.
  - Update the prior: $p(\boldsymbol{w}') = p(\boldsymbol{w}|D)$
- Beta distribution is a good choice for modelling multiple binary distribution: flipping coin example

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

  - Highlight that Beta distribution encodes prior knowledge about the parameter w before observing data.
  - The settings of $(a-1)$ and $(b-1)$ are to control **weak belief** with insufficient observation (controlled by $a + b - 2$) :
    - With 1 head and 1 tail, no solution given. ($1 + 1 - 2 = 0$) $\leftarrow$ **weak belief**
    - With 2 head and 3 tail, the prior turns out to be:

      $$\frac{\partial \ln \{p(D|a)\}}{\partial a} = \frac{2-1}{a} - \frac{3-1}{1-a} = 0 \Rightarrow 1 - a = 2a \Rightarrow a = \frac{1}{3}$$

- Hence, flipping coin (binary problem) example using Bayesian approach:
  - Collect available data

- Use Beta distibution encode prior: $p(w) \propto w^{a-1}(1-w)^{b-1}$
- Use prior to calculate the likelihood: $p(D|w) \propto w^{|H|}(1-w)^{|T|}$
- Calculate the posterior by Bayes' theorem:

  $p(w|D) \propto w^{|H|+a-1}(1-w)^{|T|+b-1}$

$\Rightarrow p(w|D) \propto Beta(w||H|+a, |T|+b)$

- Update the prior: $p(\boldsymbol{w'}) = p(\boldsymbol{w}|D)$

> Note that:
>
> If we want to predict the probability that next toss is head, we use the observation D to calculate the posterior $p(H|D)$.
>
> Since the toss result 'Head' is depend on w, we don't know it yet, we can integrate over all possible value of w:
>
> $p(H|D) = \int_0^1 p(H,w|D)dw = \int_0^1 p(H|w)p(w|D)dw$
>
> (Product rule + Conditional Independence assumption applied)
>
> If we already know w, then the result for 'Head' only depends on this knowledge. All other information is no longer important: $p(H|w,D) = p(H|w)$
>
> Then:
>
> $p(H|D) = \int_0^1 p(H,w|D)dw = \int_0^1 p(H|w)p(w|D)dw$
>
> $= \int_0^1 w^{|H|}(1-w)^{|T|}p(w|D)dw$
>
> $= wBeta(w||H|+a, |T|+b)$
>
> $= E[w|D] = \dfrac{|H|+a}{|H|+|T|+a+b}$
>
> It indicates that we predict based on the prior (a and b) and the observed data ($|H|$ and $|T|$).

Prior: Strong belief vs Weak belief

Flipping coin example: a for the number of Head, b for the number of Tail,

- $a = b = 100 \rightarrow$ strong belief the coin is fair.
- $a = b = 2 \rightarrow$ week belief that the coin is fair.
- $a = 100, b = 1 \rightarrow$ strong belief that the coin is biased towards head.