Exercise 11-1  Frequentist versus Bayesian Statistics

Consider this – rather pathological – example to illustrate the difference between frequentist and bayesian statistics: Alice and Bob play a game in which the first person to get 6 points wins. The points are scored in the following way: A referee is standing at a pool table Alice and Bob cannot see. Before the game begins, the referee rolls a ball onto the table coming to rest at a random position. Each point scored is decided by the referee rolling another ball. If the ball comes to rest left of (the middle of) the initial ball, Alice scores, if it comes to rest right, Bob scores. The players know nothing but who scored a point. If the portion left of the initial ball is denoted as $p$, it is obvious that the probability of Alice scoring a point is $p$.

Now, consider the following situation within the game: Alice has 5 points, Bob has 3. Let us investigate the probability of Bob winning.

(a) Assume that the initial ball came to rest such that $p = 2/3$. What is the probability that Bob wins?

(b) Unfortunately, we do not know $p$ – we only have some data we can try to estimate it from. Follow a frequentist approach: compute the maximum likelihood estimator for $p$ and the probability of Bob winning.

(c) Now, let us follow a bayesian approach: We know that $p$ is only dependent on the position of the initial ball which we assume to be uniformly distributed on the table, i.e., $\text{unif}[0, 1]$. Note that we compute the expected probability of Bob winning, as $p$ itself is now drawn from a distribution. Hint: You will need the beta function:

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{(x-1)!(y-1)!}{(x+y-1)!}$$

Exercise 11-2  Linear Regression with Gaussian Noise

Let $D, d_i = (x_{i,1}, \ldots, x_{i,M}, y_i)^T$, be a dataset of size $N$ with $M$ features and an output $y$ which depends linearly on $X$. Due to erroneous measurements the inputs are noisy, i.e.:

$$y_i = x_i^Tw + \epsilon_i,$$

where $\epsilon_i$ is the noise of data point $i$. Furthermore, assume $\epsilon$ to be gaussian distributed:

$$P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}\epsilon_i^2}.$$  

Given the variables $X$ and the model $w$, we can then model the distribution of $y$ as follows:

$$P(y_i|x_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i-x_i^Tw)^2}.$$
(a) Determine the parameter $\hat{w}$ which maximizes the probability of the training data $P(D|w)$, using the maximum-likelihood estimator: $\hat{w}_{ML} = \arg \max_w P(D|w)$.

You may assume that the $w$ are distributed independently of the input data $X$.

(b) A common assumption for the a priori distribution of random variables is:

$$P(w) = \frac{1}{(2\pi\alpha)^{M/2}} e\left(-\frac{1}{2\alpha^2} \sum_{j=0}^{M-1} w_j^2 \right)$$

Compute the parameter $\hat{w}$ which maximizes $P(w)P(D|w)$. Does this give an alternative interpretation to the $\lambda$-term of the penalized least squares function (PLS)?

**Exercise 11-3  Bias vs. Variance**

Consider the following learning curves showing training and cross-validation error for two different models when the respective model is trained on an increasing amount of data. For both cases, discuss bias and variance. What are indicators for a high variance or a high bias problem?

![Learning Curves](image)