Homework 1

1 *[LFD Exercise 1.2]* Suppose that we use a perceptron to detect spam messages. Let's say that each email message is represented by the frequency of occurrence of keywords, and the output is +1 if the message is considered spam.

1. Can you think of some keywords that will end up with a large positive weight into perceptron?

Answer. Some keywords for a large positive weight – Winner, Congratulations, Viagra, Rich, Free, Instant

2. How about keywords that will get a negative weight?

Answer. Keywords suggested for negative weight - Reminder, Alert, Register, Meeting

3. What parameter in the perceptron directly affects how many borderline messages end up classified as spam?

Answer. The bias term acts as the threshold in the perceptron learning algorithm. Consider the weights to be w_1, w_2, $w_3 \ldots w_d$. The boundary message will be classified as spam if the sum of weights and input values is greater than bias, i.e $\sum w_i x_i >$-b, where b is the bias term. According to the formula above, if we have a larger bias, the RHS will decrease and the boundary shifts downwards, so more borderline cases will be classified as spam. If bias decreases, then boundary shifts upwards and fewer cases will be classified as spam.

2 Consider a coin tossing experiment.

1. Given only that the true probability, θ, satisfies $0 \leq p \leq 1$, what is a best estimate of θ? (There is no additional information that you need to answer this question, but you may opt to make an assumption or experiment if it helps.)

Answer.

1) a. No prior information: we can estimate the model and assume the coin is fair, so probability is **1/2** before any experiments.
b. We conduct experiments to build a model by assuming the probability of getting a head as θ, conducting trials and then equating the probability to be number of heads over total number of tosses.

(Problem continued) We denote the probability of heads of this coin as θ. For any value of θ, the probability of k heads in n tosses is given by the Binomial distribution: $Pr(k|\theta, n) = \binom{n}{k}(\theta)^k(1-\theta)^{n-k}$. Say that 100 tosses of the same coin results in heads 70 times and tails 30 times.

2. Build a model using maximum likelihood estimation (MLE) to infer θ. Which value of θ is most likely?

Answer. Consider we have n tosses. We want to maximise the number of times heads appears. Let k represent the number of heads. The likelihood of θ as a function of X is defined as

$$\text{lik}(\theta) = \text{f(x/ } \theta) = \binom{n}{k}(\theta)^k(1-\theta)^{n-k}.$$

Rather than maximizing the actual likelihood, we maximize the log likelihood as it is easier to compute.

$$\text{l}(\theta) = \sum_{i=1}^{n} \log[\text{f}(x_i/\ \theta)] = \sum_{i=1}^{n} \log[\binom{n}{k}(\theta)^k(1-\theta)^{n-k}]$$

Discarding the combinatorial part as this is a log sum, we consider the term

$$k\log\theta+(n-k)\log(1\text{-}\theta)$$

We solve for the θ that maximizes this function by differentiating the above formula w.r.t θ. This gives us

$$\frac{k}{\theta} - \frac{n-k}{1-\theta} = 0$$

Solving this equation for θ, we get

$$\theta = \frac{k}{n}$$

Since number of heads is $k = 70$ and number of tosses is n=100, we derive

$$\theta = 0.7$$

3. Can we judge that this is an unfair coin? Explain your Answer.

Answer. Let us consider the null hypothesis H_0 that the coin is unbiased, which means that the probability of heads appearing is 0.5. The alternate hypothesis H_α is that the coin is biased and θ is not equal to 0.5. We choose to construct a 95% confidence interval with the significance level α as 0.05. Using the formula

$$\mu \pm \text{Z}\frac{s}{\sqrt{n}}$$

our confidence interval is $[0.40, 0.60]$. Our value of 0.7 does not lie within this interval, so we reject the null hypothesis.
Suppose that the null hypothesis is true. The number of heads will have binomial distribution with mean 100*0.5 = 50 and standard deviation $\sqrt{(100)(1/2)(1/2)} = 5$. In our experiment, we have 20 more heads. The probability that in tossing a fair coin the number of heads differs from 50 is

$$\sum_{k=70}^{100} \binom{100}{k} (0.50)^{100}$$

which gives us 0.0000196. At the 95% confidence level, we can reject the null hypothesis that this coin is unbiased.

3 In the programming logistic regression, part (c), compare our round-robin version of gradient descent which deterministically uses the next point in turn to perform the gradient descent, versus the standard, stochastic form which chooses a single point at random for an iteration. Describe whether you think this is a good robust idea or not for datasets in general.

Answer. The round robin method is not a robust idea. If our dataset is not even and has a lot of valleys and peaks, then this is not a good idea as there's a high chance of getting stuck in a local minima. In this case, picking points at random using Stochastic Gradient Descent would be better.
Another issue would be if the number of iterations is a lot smaller than the training data size, in which case the entire dataset is not utilized and we may end up with weights that are biased. SGD is again better in this case as all points have an equal probability of being picked.