HOMEWORK 2
KERNEL SVM AND PERCEPTRON

DANA VAN AKEN

Problem 2: Understanding the Likelihood Function, Bit by Bit

1. (a) $H_i \sim Bernoulli(p_H), N_H \sim Binomial(N_{bits}, p_H)$.

   (b) i. $L(H_1, ..., H_{Nbits}; p_H) = \prod_{i=1}^{N_{bits}} P(H_i; p_H) = \prod_{i=1}^{N_H} p_H \prod_{i=1}^{N_{bits}-N_H}(1-p_H) = p_H^{N_H}(1-p_H)^{N_{bits}-N_H}$.

   ii.

   iii. [000]: $\int_0^1 p_H^0 (1-p_H)^3 dp_H = \int_0^1 (1-p_H)^3 dp_H = (-1)\frac{1}{4}(1-p_H)^4\big|_0^1 = \frac{-1}{4}\big((1-1)^4 - (1-0)^4\big) = \frac{1}{4}$

   [111]: $\int_0^1 p_H^3 (1-p_H)^0 dp_H = \int_0^1 p_H^3 dp_H = \frac{1}{4}p_H^4\big|_0^1 = \frac{1}{4}(1^4 - 0^4) = \frac{1}{4}$

   [110]: $\int_0^1 p_H^2 (1-p_H)^1 dp_H = \int_0^1 (p_H^2 - p_H^3) dp_H = (\frac{1}{3}p_H^3 - \frac{1}{4}p_H^4)\big|_0^1 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$

   [001]: $\int_0^1 p_H^1 (1-p_H)^2 dp_H = \int_0^1 p_H(1-p_H^2) dp_H = \int_0^1 p_H(1 - 2p_H + p_H^2) dp_H$

   $= \int_0^1 (p_H - 2p_H^2 + p_H^3) dp_H = (\frac{1}{2}p_H^2 - \frac{2}{3}p_H^3 + \frac{1}{4}p_H^4)\big|_0^1 = \frac{1}{2} - \frac{2}{3} + \frac{1}{4} = \frac{1}{12}$

   You can tell that this is not a valid probability distribution over p_H because the total sum of the area under these curves is not equal to 1. The reason that it's invalid is because these 4 sequences are only a subset of the total possible sequences, (for example, we are missing [101], [100], etc.).

   (c) We choose p_H that maximizes the probability of the observed sequence. We find this value of p_H by maximizing the likelihood function.

   $$\hat{p}_H = \arg\max_{p_H} P(H_1, ..., H_{N_{bits}}; p_H) = \arg\max_{p_H}\big(p_H^{N_H}(1-p_H)^{N_{bits}-N_H}\big)$$

   There are different techniques that can be used to find the value of p_H that maximizes the likelihood function (e.g. taking the derivative, gradient descent). To find this maximizing value, we can take the derivative of this likelihood function, set it equal to 0, and solve for p_H since a closed-form solution exists for this particular likelihood function.

   Instead of maximizing the likelihood function, we will instead maximize the log-likelihood function which reaches its maximum value at the same points as the original function, (this is because the logaritm function is monotonically increasing).

   $$\hat{p}_H = \arg\max_{p_H} log\big(P(H_1, ..., H_{N_{bits}}; p_H)\big) = \arg\max_{p_H} log\big(\big(p_H^{N_H}(1-p_H)^{N_{bits}-N_H}\big)\big)$$
   $$= \arg\max_{p_H} log(p_H^{N_H}) + log(1-p_H)^{N_{bits}-N_H} = \arg\max_{p_H} N_H log(p_H) + (N_{bits} - N_H)log(1-p_H)$$

   Now take the derivative of the log-likelihood function:

   $$\frac{d}{dp_H}\big(N_H log(p_H) + (N_{bits} - N_H)log(1-p_H)\big) = \frac{N_H}{p_H} - \frac{N_{bits}-N_H}{1-p_H}$$

   Set it equal to 0 and solve for p_H:

   $$\frac{N_H}{p_H} - \frac{N_{bits}-N_H}{1-p_H} = 0$$

   $$\frac{N_H}{p_H} = \frac{N_{bits}-N_H}{1-p_H}$$

$$\frac{N_H(1-p_H)}{p_H} = N_{bits} - N_H$$

$$\frac{1-p_H}{p_H} = \frac{N_{bits} - N_H}{N_H}$$

$$\frac{1}{p_H} - 1 = \frac{N_{bits}}{N_H} - 1$$

$$p_H = \frac{N_H}{N_{bits}}$$

This result shows that the maximizing value of p_H is the number of heads (or 1-bits) divided by the total number of flips (or bits).

2.