CS542 A1 Machine Learning
Problem Set 1
Solutions to Written Problems

4.1 Maximum Likelihood Estimate for Coin Toss

The probability distribution of a single binary variable $x \in \{0,1\}$ that takes value 1 with probability μ is given by the *Bernoulli* distribution

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

For example, we can use it to model the probability of seeing 'heads' $(\mathrm{x} = 1)$ or 'tails' $(\mathrm{x} = 0)$ after tossing a coin, with μ being the probability of seeing 'heads'. Suppose we have a dataset of independent coin flips $D = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ and we would like to estimate μ using Maximum Likelihood. Recall that we can write down the likelihood function as

$$p(D|\mu) = \prod_{i=1}^{m} p(x^{(i)}|\mu) = \prod_{i=1}^{m} \mu^{x^{(i)}}(1-\mu)^{1-x^{(i)}}$$

The log of the likelihood function is

$$\ln p(D|\mu) = \sum_{i=1}^{m} x^{(i)} \ln \mu + \left(1 - x^{(i)}\right) \ln(1-\mu)$$

Show that the ML solution for μ is given by $\mu_{ML} = \frac{h}{m}$, where h is the total number of 'heads' in the dataset. Show all of your steps.

Answer: The Maximum Likelihood solution is obtained by finding the value μ_{ML} that maximizes the likelihood function, which is the solution to

$$\frac{d}{d\mu}\left(\sum_{i=1}^{m} x^{(i)} \ln \mu + \left(1 - x^{(i)}\right) \ln(1-\mu)\right) = 0$$

Taking the derivative, we get

$$\sum_{i=1}^{m} x^{(i)} \frac{1}{\mu} + (-1)\left(1 - x^{(i)}\right) \frac{1}{1-\mu} = 0,$$

$$\sum_{i=1}^{m} x^{(i)} \frac{1}{\mu} = \sum_{i=1}^{m} \left(1 - x^{(i)}\right) \frac{1}{1-\mu},$$

$$\frac{h}{\mu} = \frac{m-h}{1-\mu}, \qquad \mu = \frac{h}{m}$$

4.2 Localized linear regression

Suppose we want to estimate *localized* linear regression by weighting the contribution of the data points by their distance to the query point $x^{(q)}$, i.e. using the cost

$$E\left(x^{(q)}\right) = \frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))^2(x^{(i)} - x^{(q)})^{-2}$$

where $(x^{(i)} - x^{(q)})^{-2} = (w^{(i)})^2$ is the inverse Euclidean distance between the training point $x^{(i)}$ and query (test) point $x^{(q)}$.

Derive the modified normal equations for the above cost function $E\left(x^{(q)}\right)$. *(Hint: first, re-write the cost function in matrix/vector notation, using a diagonal matrix to represent the weights $w^{(i)}$).*

Answer: First, re-write the cost as

$$E\left(\boldsymbol{x^{(q)}}\right) = \frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)} - \boldsymbol{\theta^T x^{(i)}}\right)^2\left(w^{(i)}\right)^2 = \frac{1}{2}\sum_{i=1}^{m}\left(w^{(i)}y^{(i)} - w^{(i)}\boldsymbol{\theta^T x^{(i)}}\right)^2$$

*Note that, unlike the SSD cost, this cost depends on the query example $\boldsymbol{x^{(q)}}$. Let **W** be a diagonal m x m matrix with diagonal entries $w^{(i)}, i = 1, \ldots, m$. Let us re-write the above cost equation in terms of matrices and vectors. Notice that multiplying the output vector $\boldsymbol{Y} = [y^{(1)} \quad \ldots \quad y^{(m)}]^{\boldsymbol{T}}$ by **W** results in the vector of weighted outputs,*

$$\boldsymbol{WY} = [w^{(1)}y^{(1)} \quad \ldots \quad w^{(m)}y^{(m)}]^{\boldsymbol{T}}$$

*and, similarly, multiplying the design matrix by **W** produces **WX**, the weighted design matrix. Replacing the original **Y** and **X** in the SSD cost with their weighted versions, we can write the weighted cost function (showing vectors/matrices in boldface for clarity) as:*

$$E\left(\boldsymbol{x^{(q)}}\right) = \frac{1}{2}\sum_{i=1}^{m}(w^{(i)}y^{(i)} - w^{(i)}\boldsymbol{\theta^T x^{(i)}})^2 = (\boldsymbol{WY} - \boldsymbol{WX\theta})^2$$

Substituting $\boldsymbol{Y} = \boldsymbol{WY}$ and $\boldsymbol{X} = \boldsymbol{WX}$ into the normal equations, the solution is

$$\boldsymbol{\theta_{ML}} = \left((\boldsymbol{WX})^{\boldsymbol{T}}(\boldsymbol{WX})\right)^{-1}(\boldsymbol{WX})^{\boldsymbol{T}}(\boldsymbol{WY})$$

4.3 Betting on Trick Coins

A game is played with three coins in a jar: one is a normal coin, one has “heads” on both sides, one has “tails” on both sides. All coins are “fair”, i.e. have equal probability of landing on either side. Suppose one coin is picked randomly from the jar and tossed, and lands with “heads” on top. What is the probability that the bottom side is also “heads”? Show all your steps.

Answer: At first, it may seem that the answer is one-half: We know that the tails/tails coin has not been picked, and only one of the remaining two—the heads/heads coin—can have the down-side be heads. However, the correct answer is two thirds, so you should not bet on the intuitive estimate! Let the following variables designate possible outcomes:

BH: bottom side of picked coin is heads
TH: top side of picked coin is heads
H : heads/heads coin is picked, 1/3
T : tails/tails coin is picked, 1/3
N : heads/tails coin is picked, 1/3

Then, using rules of probability

$$p(BH|TH) = \frac{p(BH,TH)}{p(TH)} =$$

$$\frac{p(BH,TH)}{p(TH|H)p(H) + p(TH|T)p(T) + p(TH|N)p(N)} = \frac{\frac{1}{3}}{1 * \frac{1}{3} + 0 * \frac{1}{3} + \frac{1}{2} * \frac{1}{3}} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$