CS 534: Machine Learning
Assignment 1

Sanad Saha
OSU ID: 933 620 612

06 October 2018

1. (Probability) Consider two coins, one is fair and the other one has a 1/10 probability for head. Now you randomly pick one of the coins, and toss it twice. Answer the following questions.

(a) What is the probability that you picked the fair coin? What is the probability of the first toss being head?

**Answer:**

Let, F be the event that the fair coin is picked. There are 2 coins in the sample space. 1 coin is fair and other one unfair. So,

$$p(F) = \frac{1}{2} \text{ (ans.)}$$

Probability of unfair coin being picked is $p(\sim F) = \frac{1}{2}$

Let, H be the event that Head comes from the first toss. This can happen in 2 ways.

First, a fair coin was picked and result is head = $\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

Second, an unfair coin was picked and head came = $\frac{1}{2} * \frac{1}{10} = \frac{1}{20}$

So, $p(H) = \frac{1}{4} + \frac{1}{20} = \frac{3}{10}$ (ans.)

(b) If both tosses are heads, what is the probability that you have chosen the fair coin?

**Answer**:

Let, T be the event that both tosses are head. $p(F|T) = ?$

From the question we can devise that $p(T|F) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4}$ and

$$p(T|\sim F) = \frac{1}{10} * \frac{1}{10} = \frac{1}{100}$$

Using Bayes theorem,

$$p(F|T) = \frac{p(T|F)p(F)}{p(T|F)p(F) + p(T|\sim F)p(\sim F)}$$

$$= \frac{\frac{1}{4} * \frac{1}{2}}{\frac{1}{4} * \frac{1}{2} + \frac{1}{2} * \frac{1}{100}} = \frac{25}{26}$$

2. (Maximum likelihood estimation for uniform distribution.) Given a set of i.i.d. samples ( $x_{1,}\ x_{2,}\ x_{3,}\ ........x_n\ \sim uniform(0;\theta)$.

(a) Write down the likelihood function of θ.

**Answer**:

Uniform distribution follows the following function : $f(x) = \frac{1}{b-a}$ for $a \le x \le b$ and 0 otherwise. For the given question probability density function of the uniform distribution is:

$$p(x_i|\,\theta) = \frac{1}{\theta}\,for\ (\ 0 \le x_i \le \theta)$$

Likelihood of this function is:

$$L(\theta) = p(x_{1,}\ x_{2,}\ x_{3,}\ ........x_n\ ;\ \theta)$$

As $x_{1,}\ x_{2,}\ x_{3,}\ ........x_n$ is independent and identically distributed we can write the likelihood in following way:

$$L(\theta) = \prod_{i=1}^{n} p(x_{i,}\ ;\theta)$$
$$= \frac{1}{\theta^n}\ for\,(\,0 \le x_i \le \theta)$$

(b) Find the maximum likelihood estimator for θ.

**Answer**:

Let log likelihood of $L(\theta) = l(\theta) = log \prod_{i=1}^{n} p(x_{i,}\ ;\theta)$

$$= log \sum_{i=1}^{n} \frac{1}{\theta} = \sum_{i=1}^{n} log\,(\,\tfrac{1}{\theta}) = n\ log(\tfrac{1}{\theta})\ (0 \le x_i \le \theta)$$

From the above equation we can find the log likelihood and see that $l(\theta)$ increases as $\theta$ decreases. To achieve maximum likelihood we need to find the maximum value of $l(\theta)$. Now, this can be achieve by keeping the distribution as tight as possible capturing all the data points. According to this equation $\theta$ can be decreased as long as $\theta$ doesn't reach the max value of $x_i$. So maximum likelihood can be reached for, $\theta = max\,(x_i)\ where\ (i = 1\ to\ n)$

3. In class when discussing linear regression, we assume that the Gaussian noise is independently identically distributed. Now we assume the noises $\varepsilon_1,\ \varepsilon_2,\ ....,\ \varepsilon_n$ are independent but each $\varepsilon_m \sim N(0,\ \sigma_m^2)$, i.e., it has its own distinct variance.

(a) Write down the log likelihood function of w.

**Answer:**

As, $\varepsilon_m \sim N(0,\ \sigma_m^2)$, $p(\varepsilon_m) = \frac{1}{\sigma\sqrt{2\pi}}\ exp(-\ \frac{\varepsilon_m^2}{\sigma_m^2})$

**So,** $p(y_i|\,x_i;\ w) = \frac{1}{\sigma_i\sqrt{2\pi}}\ exp(\ -\ \frac{(y_i - w^T x_i)^2}{2\sigma_i^2})$

Weighted log likelihood is the following:

$$l(w) = log \sum_{i=1}^{n} p(y_i | x_i;\ w) = log \sum_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} exp(-\frac{(y_i - w^T x_i)^2}{2\sigma_i^2})$$

$$= \sum_{i=1}^{n} [\ log \frac{1}{\sigma_i\sqrt{2\pi}} + log\ (\ exp(-\frac{(y_i - w^T x_i)^2}{2\sigma_i^2}))\ ]$$

$$= \sum_{i=1}^{n} [\ log \frac{1}{\sigma_i\sqrt{2\pi}} - \frac{(y_i - w^T x_i)^2}{2\sigma_i^2}\ ]$$

$$= \sum_{i=1}^{n} log \frac{1}{\sigma_i\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2\sigma_i^2}$$

$$= \sum_{i=1}^{n} log \frac{1}{\sigma_i\sqrt{2\pi}} - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (y_i - w^T x_i)^2$$

$$= \sum_{i=1}^{n} log \frac{1}{\sigma_i\sqrt{2\pi}} + \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (w^T x_i - y_i)^2$$

(b) Show that maximizing the log likelihood is equivalent to minimizing a weighted least square loss function $J(\mathbf{W}) = \frac{1}{2}\sum_{m=1}^{n} a_m(\mathbf{w}^T\mathbf{x}_m - y_m)^2$, and express each $a_m$ in terms of $\sigma_m$.

**Answer**:

From the log likelihood function above it can be seen that the maximum of log likelihood depends on minimizing the term $\sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (w^T x_i - y_i)^2$. So, the function for maximizing the log likelihood can be written as follows:

$$j(w) = \sum_{i=1}^{n} \frac{1}{2\sigma_i^2}(w^T x_i - y_i)^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (w^T x_i - y_i)^2$$

$$= \frac{1}{2} \sum_{i=1}^{n} a_i\ (w^T x_i - y_i)^2\ [\ a_i = \frac{1}{\sigma_i^2}] \text{........(1)}$$

According to the question numbering all the data by m instead of i, where m = 1...n we can turn the equation into the following:

$$\text{j(w)} = \frac{1}{2} \sum_{m=1}^{n} a_m\ (w^T x_m - y_m)^2\ [\ a_m = \frac{1}{\sigma_m^2}]$$

(proved)

(c) Derive a batch gradient descent algorithm for optimizing this objective.
**Answer**:
First we have to calculate the gradient of the maximum log likelihood function. Which is shown below:

$$j(w) = \frac{1}{2} \sum_{i=1}^{n} a_i (w^T x_i - y_i)^2$$

$$\text{Gradient}(j(w)) = \frac{\partial j(w)}{\partial w} = \frac{\partial}{\partial w} \frac{1}{2} \sum_{i=1}^{n} a_i (w^T x_i - y_i)^2$$

$$= \sum_{i=1}^{n} a_i (w^T x_i - y_i) \frac{\partial}{\partial w} (w^T x_i - y_i)$$

$$= \sum_{i=1}^{n} a_i (w^T x_i - y_i) x_i$$

Batch gradient descent:
Given training example $(x_i, y_i)$ $i = 1, ...., N$
Let $w = w_0$
Repeat until convergence:
For i = 1 to N

$$w = w - \lambda * a_i (w^T x_i - y_i) x_i$$

(ans.)

(d) Derive a closed form solution to this optimization problem
**Answer**:
The log likelihood is:

$$j(w) = \frac{1}{2} \sum_{i=1}^{n} a_i (w^T x_i - y_i)^2$$

Now let X be the matrix of $x_i$ , y is a vector where y = transpose$\{y_1, y_2, ....., y_i\}$
A is a vector consisting of $a_i$. so the log likelihood can be written as:

$$J(w) = (y - Xw)^T A (y - Xw)$$

$$= (y - Xw)^T (Ay - AXw)$$

$$= (y^T - X^T w^T) (Ay - AXw)$$

$$= (y^T - X^T w^T) (Ay - AXw)$$

$$= y^T Ay - 2 X^T w^T Ay + - X^T w^T Axw$$

Now, for maximum log likelihood:

$\frac{\partial}{\partial w} j(w) = \frac{\partial}{\partial w}(y^T Ay - 2X^T w^T Ay + - X^T w^T Axw)$

$\frac{\partial}{\partial w}(y^T Ay) = 0.$

So, $\frac{\partial}{\partial w} j(w) = \frac{\partial}{\partial w}(-2X^T w^T Ay + - X^T w^T Axw)$

Now setting the $\frac{\partial}{\partial w} j(w) = 0,$

$\frac{\partial}{\partial w}(-2X^T w^T Ay + - X^T w^T Axw) = 0.$

So Finding the $\frac{\partial}{\partial w}(-2X^T w^T Ay + - X^T w^T Axw)$ will give us the closed form solution.

4. (Decision theory). Consider a binary classification task with the following loss matrix:

| predicted label $\hat{y}$ | true label $y$ = 0 | true label $y$ = 1 |
|---|---|---|
| 0 | 0 | 10 |
| 1 | 5 | 0 |

We have build a probabilistic model that for each example $x$ gives us an estimated $P(y = 1|x)$. It can be shown that, to minimize the expected loss for our decision, we should set a probability threshold $\theta$ and predict $\hat{y} = 1$ if $P(y = 1|x) > \theta$ and $\hat{y} = 0$ otherwise.

(a) Please compute the θ for the above given loss matrix.

**Answer:**

We want to predict $\widehat{y} = 1$ *if* $p(y = 1|x) > \theta$.

Given loss matrix $L(y, \widehat{y})$ the estimated loss of predicting $\widehat{y} = 1$ *is*

$$p(y = 0|x) * L(0,1) = p(y = 0|x) * 5$$

And the estimated loss of predicting $\widehat{y} = 0$ *is*

$$p(y = 1|x) * L(1,0) = p(y = 1|x) * 10$$

Now, we can predict $\widehat{y} = 1$ if,

$$p(y = 0|x) * 5 < p(y = 1|x) * 10 \ldots\ldots\ldots (1)$$

Let, z = $p(y = 1|x)$

So, $p(y = 0|x) = 1 - z$

So, equation 1 can be rewritten as following,

$$(1 - z) * 5 < z * 10$$

$$\approx 5 < 15z$$

$$\approx z > \frac{1}{3}$$

$$\approx p(y = 1|x) > \frac{1}{3}$$

As a result if we set our threshold $\theta = \frac{1}{3}$ then we will be able to predict $\hat{y} = 1$ if threshold is greater than $\frac{1}{3}$ and $\hat{y} = 0$ otherwise.

(b) Show a loss matrix where the threshold is 0.1

**Answer:**

The following loss matrix has the threshold of 0.1:

| Predicted label $\hat{y}$ | True label y | |
|---|---|---|
| | 0 | 1 |
| 0 | 0 | 9 |
| 1 | 1 | 0 |

5. Consider the maximum likelihood estimation problem for multi-class logistic regression using the softmax function defined below:

$$p(y = k|x) = \frac{exp(w_k^T x)}{\sum_{j=1}^{K} exp(w_j^T x)}$$

We can write the likelihood function as:

$$L(w) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y = k \mid x_i)^{I(y_i=k)}$$

Where $I(y_i = k)$ is the indicator function, taking value 1 if $y_i$ *is* $k$.

(a) What are i and k in this likelihood function?

**Answer:**

In this likelihood function i points to specific training example $(x_i, y_i)$ where i can be 1...N. And k points to any class of y as there are K possible classes.

(b) Compute the log-likelihood function

**Answer:**

Log likelihood function,

$$l(w) = log\ L(w) = log \prod_{i=1}^{N} \prod_{k=1}^{K} p(y = k \mid x_i)^{I(y_i=k)}$$

$$= log \prod_{i=1}^{N} \prod_{k=1}^{K} \left( \frac{exp(w_k^T x_i)}{\sum_{j=1}^{K} exp(w_j^T x_i)} \right)^{I(y_i=k)}$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = k) \, [\, w_k^T x_i - log \sum_{j=1}^{K} exp(w_j^T x_i) \,]$$

$$= \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = k) \, w_k^T x_i - \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = k) \, log \sum_{j=1}^{K} exp(w_j^T x_i)$$

(ans.)

(c) What is the gradient of the log-likelihood function w.r.t the weight vector $w_c$ of class c? (Precursor to this question, which terms are relevant for $w_c$ in the log likelihood function?)

**Answer:**

Both terms of log likelihood, $l(w)$ is relevant for $w_c$.

Gradient of $l(w) = \frac{\partial\, l(w)}{\partial w_c}$

$$= \frac{\partial}{\partial w_c} \left( \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = c) \, w_k^T x_i - \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = k) \, log \sum_{j=1}^{K} exp(w_j^T x_i) \right)$$

$$= \frac{\partial}{\partial w_c} \left( \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = c) \, w_k^T x_i \right) - \frac{\partial}{\partial w_c} \left( \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = k) \, log \sum_{j=1}^{K} exp(w_j^T x_i) \right)$$

$$= \prod_{i=1}^{N} I(y_i = c) \, x_i - \frac{\partial}{\partial w_c} \left( \prod_{i=1}^{N} \prod_{k=1}^{K} I(y_i = k) \, log \sum_{j=1}^{K} exp(w_j^T x_i) \right)$$

$$= \prod_{i=1}^{N} I(y_i = c) \, x_i - \prod_{i=1}^{N} \frac{\partial}{\partial w_c} \left( log \sum_{j=1}^{K} exp(w_j^T x_i) \right) \quad [\prod_{k=1}^{K} I(y_i = k) = 1]$$

$$= \prod_{i=1}^{N} I(y_i = c) \, x_i - \prod_{i=1}^{N} \frac{\partial}{\partial w_c} \left( log \sum_{j=1}^{K} exp(w_j^T x_i) \right)$$

$$= \prod_{i=1}^{N} I(y_i = c)\, x_i - \prod_{i=1}^{N} \frac{1}{\sum_{j=1}^{K} exp(w_j^T x_i)} \frac{\partial}{\partial w_c} (\ exp(w_j^T x_i))$$

$$= \prod_{i=1}^{N} I(y_i = c)\, x_i - \prod_{i=1}^{N} \frac{exp(w_c^T x_i)}{\sum_{j=1}^{K} exp(w_j^T x_i)}\, x_i$$

$$= \prod_{i=1}^{N} [\ I(y_i = c) - \frac{exp(w_c^T x_i)}{\sum_{j=1}^{K} exp(w_j^T x_i)}\ ]\, x_i$$

$$= \prod_{i=1}^{N} [I(y_i = c) - p(y = c \mid x_i)\ ]\, x_i$$

(ans.)