Written Homework 1:

Behnam Saeedi
(Saeedib@oregonstate.edu)

CS534: Machine Learning
Due Oct 6th 11:59pm
Fall 2018

1 QUESTION 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.

1.1 A

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

1.1.1 Answer

The probability would of picking the fair coin is $\frac{1}{2}$

The probability of first toss being head is $P(Head) \times P(Fair) + P(unfair) \times P(head)$

$$\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{10} = \frac{1}{4} + \frac{1}{20} = \frac{6}{20}$$
$$= \frac{3}{10}$$

1.2 B

If both tosses are heads, what is the probability that you have chosen the fair coin (Hint: Bayes Rule)?

1.2.1 Answer

let A represent a fair coin ($P(A)$ is probability off coin being fair); And B represent two heads in a row ($P(B)$ is probability of getting two heads in the row). We can use the following equation to compute $P(A|B)$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$
$$P(A) = \frac{1}{2}, P(B|A) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$
$$P(B) = P(fair) \times (P(fair - heads))^2 + P(unfair) \times (P(unfair - heads))^2$$
$$P(B) = \frac{1}{2} \times (\frac{1}{2})^2 + \frac{1}{2} \times (\frac{1}{10})^2$$
$$\Rightarrow P(A|B) = \frac{[\frac{1}{2} \times \frac{1}{2}] \times [\frac{1}{2}]}{[\frac{1}{2} \times (\frac{1}{2})^2 + \frac{1}{2} \times (\frac{1}{10})^2]}$$
$$\Rightarrow P(A|B) = \frac{1}{1 + \frac{1}{25}}$$
$$\Rightarrow P(A|B) = \frac{1}{\frac{26}{25}}$$
$$\Rightarrow P(A|B) = \frac{25}{26}$$

---

2 Question 2

(Maximum likelihood estimation for uniform distribution.) Given a set of i.i.d. samples $x_1, x_2, ..., x_n \sim$ uniform$(0, \theta)$.

2.1 A

Write down the likelihood function of θ.

2.1.1 *Answer*

According to question we know that the $P(y^i|x_i;\theta) = \frac{1}{\theta}$ for all $0 \leq x_i \leq \theta$. Given that Likelihood function $L(\theta)$ is defined as:

$$L(\theta) = \prod_{i=1}^{n} P(y^i|x^i;\theta)$$

Therefore, the maximum likelihood function is defined as:

$$\prod_{i=1}^{n} \frac{1}{\theta} \Rightarrow \theta^{-n}$$

2.2 B

Find the maximum likelihood estimator for θ.

2.2.1 *Answer*

Now we can use the likelihood function for estimator MLE $\ell(\theta)$:

$$\ell(\theta) = \frac{\partial logL(\theta)}{\partial\theta}$$

$$\Rightarrow \ell(\theta) = \frac{-n}{\theta} < 0$$

In order to make sure all x_i is bound by 0 and θ we need to find the largest x_i and set our $\hat{\theta}$ to it. Recall that $L(\theta|x)$ is always greater or equal to x_n:

$$\Rightarrow \hat{\theta} = x_n$$

Now we are certain that all $x_i \in U(0, \theta)$ and 0 everywhere else.

---

3 Question 3

(Weighted linear regression) In class when discussing linear regression, we assume that the Gaussian noise is independently identically distributed. Now we assume the noises $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are independent but each $\epsilon_m \sim N(0, \sigma_m^2)$, i.e., it has its own distinct variance.

3.1 A

Write down the log likelihood function of w.

3.1.1 Answer

$$\ell(\theta) = logL(\theta)$$

$$log \prod_{m=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(w^T x_m - y_m)^2}{2\sigma^2})$$

The difference in this derivation is the σ.

$$\sum_{m=1}^{n} log \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(w^T x_m - y_m)^2}{2\sigma^2})$$

$$nlog \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2} \sum_{m=1}^{n} \frac{1}{\sigma^2} (w^T x_m - y_m)^2$$

3.2 B

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

3.2.1 Answer

This is simple because the positive term is not changing and maximizing the whole term depends on minimizing the error term. (a_m is shown in this case as $\frac{1}{\sigma_m^2}$). because:

therefore:

$$max(nlog \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2} \sum_{m=1}^{n} \frac{1}{\sigma_m^2} (w^T x_m - y_m)^2) \Rightarrow min(-\frac{1}{2} \sum_{m=1}^{n} \frac{1}{\sigma_m^2} (w^T x_m - y_m)^2)$$

3.3 C

Derive a batch gradient descent algorithm for optimizing this objective.

3.3.1 Answer

In order to do this we need to compute the partial derivative of the $\ell(\theta)$ from above equation.

$$\frac{\partial \ell(\theta)}{\partial \ell(w^T)}$$

$$\frac{\partial \frac{1}{2} \sum_{m=1}^{n} \frac{1}{\sigma_m^2} (w^T x_m - y_m)^2}{\partial w^T} = \frac{1}{2} \sum_{m=1}^{n} \frac{1}{\sigma_m^2} 2(w^T x_m - y_m) x_m$$

3.4 D

Derive a closed form solution to this optimization problem.

3.4.1 *Answer*

Let X be an $m \times n$ matrix containing the training set and y be an m dimensional vector with target values. We know that:

$$h_w(x_i) = (x_i)^T w \Rightarrow Xw - y = H$$

$$\Rightarrow \frac{1}{2}(Xw - y)^T(Xw - y) = J(w)$$

$$\Rightarrow \nabla_w J(w) = \nabla_w \frac{1}{2}(Xw - y)^T(Xw - y)$$

$$\Rightarrow \frac{1}{2}\nabla_w (W^T X^T Xw - W^T X^T y - y^T XW + y^T y)$$

$$\Rightarrow \nabla_w J(w) = \frac{1}{2}(X^T XW + X^T XW - 2X^T y)$$

$$\Rightarrow \nabla_w J(w) = X^T XW - X^T y = 0$$

$$\Rightarrow X^T XW = X^T y$$

$$\Rightarrow W = (X^T X)^{-1} X^T y$$

Please note this is the estimated coefficient (slope) and a line equation could be devised to estimate y:

$$\hat{y} = X(X^T X)^{-1} X^T y$$

---

4 QUESTION 4

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

	True label y		
Predicted label $\hat{y}$		0	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 and predict $\hat{y} = 1$ if $P(y = 1|x) > \theta$ and $\hat{y} = 0$ otherwise.

4.1 A

Please compute the θ for the above given loss matrix.

4.1.1 Answer

This is fairly strait forward. We are trying to minimize loss. In this case we ignore the cases which we are correct since such cases have no loss. This simplifies the table to the following:

True label y	
0	1
5	10

Now we can consider that we are trying to minimize the amount of loss. We can assume that we are wrong regardless the decision we make. In that case it is beneficial for us to set the thresh hold in such way that in case that we are wrong we are going to lose less. This optimization look as following:

$$P(y = 0|x)5 \leq P(y = 1|x)10$$

We know that $P(y = 0|x) = 1 - P(y = 1|x)$ so let's refer to $P(y = 0|x)$ as P. so we can plug that in the above inequality:

$$(1 - P)5 \leq (P)10 \Rightarrow 5 - 5P \leq 10P$$

$$\Rightarrow -15P \leq -5 \Rightarrow P \geq \frac{5}{15}$$

$$\Rightarrow P \geq \frac{1}{3}$$

in order to optimize θ we need to minimize P and the smallest value we can get for it is $\frac{1}{3}$.

$$\Rightarrow \theta = \frac{1}{3}$$

4.2 B

Show a loss matrix where the threshold is 0.1.

4.2.1 Answer

in order to get that thresh hold the we need $P \geq \frac{1}{10}$. This could be achieved by $1P \leq 9P$. We can use the matrix:

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

5 QUESTION 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 out the likelihood function as:

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

5.1 A

What are i and k in this likelihood function?

5.1.1 *Answer*

- The variable i is the training example index. the range of is is [1,N]
- K indicates the classification labels for each class.

5.2 B

Compute the log-likelihood function.

5.2.1 *Answer*

Likelihood function could be drived using the following precedure:

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

$$\ell(w)\ is\ defined\ as:\ log\ L(w)$$

$$\Rightarrow \ell(w) = log \prod_{i=1}^{N} \prod_{k=1}^{K} p(y=k|x_i)^{I(y_i=k)}$$

we can move the first product out of log and turn it into a sum

$$\Rightarrow \ell(w) = \sum_{i=1}^{N} log \prod_{k=1}^{K} p(y=k|x_i)^{I(y_i=k)}$$

$$p(y=k|x_i) = [\frac{exp(w_k^T x_i)}{\sum_{j=1}^{k} exp(w_k^T x_i)}]$$

$$\Rightarrow \ell(w) = \sum_{i=1}^{N} log \prod_{k=1}^{K} [\frac{exp(w_k^T x_i)}{\sum_{j=1}^{k} exp(w_k^T x_i)}]^{I(y_i=k)}$$

now we can move the second product out of the log:

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

5.3 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?)

5.3.1 Answer

... sweet lord. As we computed log-likelihood function in last part (pray to lord it is correct) we can say:

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

we can expand this equation further by multiplying the outside of bracket portion in:

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

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

we can compute the partial of the last portion to be:

$$\frac{\partial \sum_{j=1}^{K} exp(w_j^T x_i)}{\partial w_k} = \sum_{j=1}^{k} x_i exp(w_j^T x_i)$$

$$\Rightarrow \frac{\partial\, \ell(w)}{\partial\, w_k} = \sum_{i=1}^{N} I(y_i = k) x_i - \sum_{i=1}^{N} I(y_i = k)\ log\ \sum_{j=1}^{k} x_i exp(w_j^T x_i)$$