AI534 — Written Homework Assignment 1 (40 pts)
Due Oct 15th 11:59pm, 2021

1. (Weighted linear regression) (15 pts) In class when discussing linear regression, we assume that the Gaussian noise is independently identically distributed. Now we assume the noises $\epsilon_1, \epsilon_2, \cdots \epsilon_n$ are independent but each $\epsilon_i \sim N(0, \sigma_i^2)$, i.e., it has its own distinct variance.

   (a) (3pts) Write down the log likelihood function of $\mathbf{w}$.

$$l(\mathbf{w}) = \log p(D|M) = \sum_{i=1}^{n} \log N(y_i|\mathbf{x}_i^T\mathbf{w}, \sigma_i^2) \quad (1)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{1}{2\sigma_i^2}(y_i - \mathbf{x}_i^T\mathbf{w})^2} \quad (2)$$

$$= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma_i} - \sum_{i=1}^{N} \frac{1}{2\sigma_i^2}(y_i - \mathbf{x}_i^T\mathbf{w})^2 \quad (3)$$

   (b) (4pts) Show that maximizing the log likelihood is equivalent to minimizing a weighted least square loss function $J(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{n} a_i(\mathbf{w}^T\mathbf{x}_i - y_i)^2$, and express each a_i in terms of σ_i.
   Maximizing the log likelihood is equivalent to minimizing the second term in the above equation, which can be represented as:

$$J(\mathbf{w}) = \frac{1}{2} \sum_{i=1,\ldots n} a_i(y_i - \mathbf{x}_i^T\mathbf{w})^2 \quad (4)$$

   where $a_i = \frac{1}{\sigma_i^2}$.

   (c) (4 pts) Derive a batch gradient descent update rule for optimizing this objective.
   Take the gradient of J (using equation 4):

$$\nabla J(\mathbf{w}) = \sum_{i=1}^{n} a_i(y_i - \mathbf{x}_i^T\mathbf{w})\mathbf{x}_i \quad (5)$$

   The update rule is as follows:

$$\mathbf{w} \leftarrow \mathbf{w} - \lambda \sum_i a_i(y_i - \mathbf{x}_i^T\mathbf{w})\mathbf{x}_i$$

   As can be seen from this solution, this is equivalent to weighting each instance differently according to the noise variance for each instance. Instances with larger noise variance are less reliable (due to larger noise), and thus contributes less (due to smaller weight) in the learning process, which is intuitive.

   (d) (4 pts) Derive a closed form solution to this optimization problem.
   Let $\mathbf{y} = \{y_1, y_2, \cdots, y_n\}^T$ be the vector storing all the y values of the training examples, and $\mathbf{X}$ be the data matrix whose rows correspond to training examples, and A be a diagonal matrix with $A(i,i) = a_i$. The objective can written in the following matrix form:

$$J(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^T A(\mathbf{y} - \mathbf{X}\mathbf{w})$$

   Take the gradient of J in vector form and set it to zero:

$$\nabla J(\mathbf{w}) = \mathbf{X}^T A(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0$$

$$2\mathbf{X}^T A\mathbf{X}\mathbf{w} = 2\mathbf{X}^T A\mathbf{y}$$

$$\mathbf{w}^* = (\mathbf{X}^T A\mathbf{X})^{-1}\mathbf{X}^T A\mathbf{y}$$

2. (14 pts) Consider the maximum likelihood estimation problem for multi-class logistic regression using the soft-max function defined below:

$$p(y = k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x})}$$

We can write out the likelihood function as:

$$L(\mathbf{w}) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y = k|\mathbf{x}_i)^{I(y_i=k)}$$

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

(a) (2 pts) Compute the log-likelihood function. *Take the log of the likelihood function, we have:*

$$l(\mathbf{w}) = \sum_{i=1}^{N} \sum_{k=1}^{K} I(y_i = k) \log p(y_i = k|\mathbf{x}_i) = \sum_{k=1}^{K} \sum_{y_i=k} \log p(y_i = k|\mathbf{x}_i)$$

(b) (12 pts) Compute the gradient of the log-likelihood function w.r.t the weight vector $\mathbf{w}_c$ of class c. (Precursor to this question, which terms are relevant for $\mathbf{w}_c$ in the loglikelihood function? Also hint: Logistic regression slide provides the solution to this problem, just need to fill in what is missing in between.)

We want to take partial gradient with respect to $\mathbf{w}_c$. Before doing that, let's break l into two parts:

$$l(\mathbf{w}) = \sum_{k \neq c}^{K} \sum_{y_i=k} \log p(y_i = k|\mathbf{x}_i) + \sum_{y_i=c} \log p(y_i = c|\mathbf{x}_i)$$

Note that we have

$$p(y_i = k|\mathbf{x}_i) = \frac{\exp \mathbf{w}_k \cdot \mathbf{x}_i}{\sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i)}$$

Take the log:

$$\log p(y_i = k|\mathbf{x}_i) = \mathbf{w}_k \cdot \mathbf{x}_i - \log(\sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i))$$

Let $z_i = \sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i)$, we have:

$$\log p(y_i = k|\mathbf{x}_i) = \mathbf{w}_k \cdot \mathbf{x}_i - \log z_i$$

Plog this into l, we have:

$$l(\mathbf{w}) = \sum_{k \neq c}^{K} \sum_{y_i=k} (\mathbf{w}_k \cdot \mathbf{x}_i - \log z_i) + \sum_{y_i=c} (\mathbf{w}_c \cdot \mathbf{x}_i - \log z_i)$$

Now take the partial gradient:

$$\frac{\partial}{\partial \mathbf{w}_c} l = -\sum_{k \neq c}^{K} \sum_{y_i=k} \frac{1}{z_i} \frac{\partial z_i}{\partial \mathbf{w}_c} + \sum_{y_i=c} \left(\mathbf{x}_i - \frac{1}{z_i} \frac{\partial z_i}{\partial \mathbf{w}_c} \right)$$

$$= \sum_{y_i=c} \mathbf{x}_i - \sum_{k=1}^{K} \sum_{y_i=k} \frac{1}{z_i} \frac{\partial z_i}{\partial \mathbf{w}_c}$$

Note that the second double summation can be simplified to $\sum_{i=1}^{N}$, where N is the total number of points. We now plug in

$$\frac{\partial z_i}{\partial \mathbf{w}_c} = \mathbf{x}_i \exp(\mathbf{w}_c \cdot \mathbf{x}_i)$$

and $z_i = \sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i)$, we have:

$$\frac{\partial}{\partial \mathbf{w}_c} l = \sum_{y_i=c} \mathbf{x}_i - \sum_{i=1}^{N} \frac{\exp(\mathbf{w}_c \cdot \mathbf{x}_i)}{\sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i)} \mathbf{x}_i$$

We will use $\hat{y}_{ic}$ (intuitively meaning the probability of instance i belong to class c) to denote

$$\frac{\exp(\mathbf{w}_c \cdot \mathbf{x}_i)}{\sum_{j=1}^{K} \exp(\mathbf{w}_j \cdot \mathbf{x}_i)}$$

and use y_{ic} to denote a binary indicator variable such that $y_{ic} = 1$ if $y_i = c$ and 0 otherwise.

Putting these new notations to use, we arrive at:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{w}_c} l &= \sum_{i=1}^{N} y_{ic}\mathbf{x}_i - \sum_{i=1}^{N} \hat{y}_{ic}\mathbf{x}_i \\ &= \sum_{i=1}^{N} (y_{ic} - \hat{y}_{ic})\mathbf{x}_i\end{aligned}$$

Therefore, a gradient ascent update rule will be:

$$\forall c \in \{1, \cdots, k\}\text{: } \mathbf{w}_c \leftarrow \mathbf{w}_c + \lambda \sum_{i=1}^{N} (y_{ic} - \hat{y}_{ic})\mathbf{x}_i$$

3. (11 pts) (Maximum A Posterior Estimation.) Suppose we observe the values of n IID random variables $X_1, \ldots, X_n$ drawn from a single Bernoulli distribution with parameter θ. In other words, for each X_i, we know that $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$. In the Bayesian framework, we treat θ as a random variable, and use a prior probability distribution over θ to express our prior knowledge/preference about θ. In this framework, $X_1, \ldots, X_n$ can be viewed as generated by:

   - First, the value of θ is drawn from a given prior probability distribution
   - Second, $X_1, \ldots, X_n$ are drawn independently from a Bernoulli distribution with this θ value.

   In this setting, Maximum A Posterior (MAP) estimation is a natural way to estimate the value of θ by choosing the most probable value given both its prior distribution and the observed data $X_1, \ldots, X_n$. Specifically, the MAP estimation of θ is given by

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname*{argmax}_{\hat{\theta}} P(\theta = \hat{\theta} | X_1, \ldots, X_n) \\ &= \operatorname*{argmax}_{\hat{\theta}} P(X_1, \ldots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta}) \\ &= \operatorname*{argmax}_{\hat{\theta}} L(\hat{\theta}) p(\hat{\theta})\end{aligned}$$

   where $L(\hat{\theta})$ is the data likelihood function and $p(\hat{\theta})$ is the density function of the prior. Now consider using a beta distribution for prior: $\theta \sim Beta(\alpha, \beta)$, whose PDF function is

$$p(\hat{\theta}) = \frac{\hat{\theta}^{(\alpha-1)}(1-\hat{\theta})^{(\beta-1)}}{B(\alpha, \beta)}$$

   where $B(\alpha, \beta)$ is a normalizing constant to make it a proper probability density function.

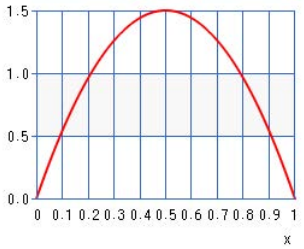
(a) (5 pts) Derive the posterior distribution $p(\hat{\theta}|X_1, \ldots, X_n, \alpha, \beta)$ and show that it is also a Beta distribution.

$$\begin{aligned} p(\hat{\theta}|X_1, \ldots, X_n) &\propto P(X_1, \ldots, X_n|\theta = \hat{\theta})p(\hat{\theta}) \\ &\propto \prod_{i=1}^{n} \hat{\theta}^{x_i}(1-\hat{\theta})^{(1-x_i)}p(\hat{\theta}) \\ &\propto \hat{\theta}^{\sum_{i=1}^n x_i}(1-\hat{\theta})^{\sum_{i=1}^n (1-x_i)}p(\hat{\theta}) \\ &\propto \hat{\theta}^{\sum_{i=1}^n x_i}(1-\hat{\theta})^{\sum_{i=1}^n (1-x_i)}\hat{\theta}^{(\alpha-1)}(1-\hat{\theta})^{(\beta-1)} \\ &\propto \hat{\theta}^{(\alpha+\sum_{i=1}^n x_i-1)}(1-\hat{\theta})^{(\beta+\sum_{i=1}^n (1-x_i)-1)} \end{aligned}$$

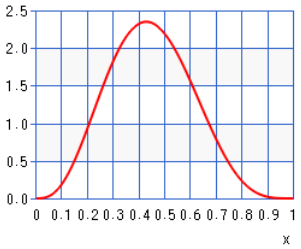
To make this a valid PDF function, we simply need to use $B\left(\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n (1-x_i)\right)$ as the denominator. So the posterior $\sim Beta(\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n (1-x_i))$.

(b) (6 pts) Suppose we use $Beta(2,2)$ as the prior, What is the posterior distribution of θ after we observe 5 coin tosses and 2 of them are head? What is the posterior distribution of θ after we observe 50 coin tosses and 20 of them are head? Plot the pdf function of these two posterior distributions. Assume that $\theta = 0.4$ is the true probability, as we observe more and more coin tosses from this coin, what do you expect to happen to the posterior?

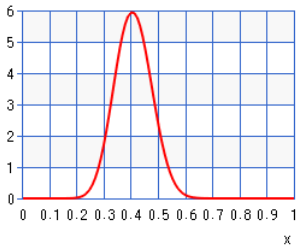
With prior $Beta(2,2)$ and an observation of n_1 heads and n_0 tails, the posterior for θ is $Beta(2+n_1, 2+n_0)$. So after observing 5 coin tosses with 2 heads, the posterior of θ becomes $Beta(4,5)$. With 50 tosses and 20 heads, the posterior becomes $Beta(22,32)$. You can see the pdf functions of the prior, and the two posteriors as follows.



(a) $Beta(2,2)$



(b) $Beta(4,5)$



(c) $Beta(22,32)$

As can be seen from the figure, the posterior becomes more and more peaked as we increase the observations. Eventually, all of the probability will concentrate at the true θ value. This is one nice property about the Bayesian approach: when we have very little data, we can fall back onto the prior to avoid catastrophic choices, and as we have more and more data they start to take over and the influence of the prior becomes increasingly neglectable.