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 $\mathbf{w}$.

   (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.

   (c) (4 pts) Derive a batch gradient descent update rule for optimizing this objective.

   (d) (4 pts) Derive a closed form solution to this optimization problem.

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.

   (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.)

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: θ $Beta(\alpha, \beta)$, 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.

(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?