AI534 — Written Homework Assignment 1 (35 pts + 8 bonus)

This written assignment covers the contents of linear regression and logistic regression. The key concepts covered here include:

- Maximum likelihood estimation (MLE)
- Gradient descent learning
- Decision theory for probabilistic classifiers
- Maximum A Posteriori (MAP) parameter estimation
- Perceptron

1. (MLE for uniform distribution) (3 pts) Given a set of i.i.d. observed samples $x_1, x_2, ..., x_n \sim \text{uniform}(0, \theta)$, we wish to estimate the parameter θ.

   (a) (1 pt) Write down the likelihood function of θ.

   (b) (2 pts) Derive the maximum likelihood estimation for θ, which is the value for *theta* that maximizes the function of part (a). (Hint: for the likelihood function is a monotonic function. So the maximizing solution is at the extreme, no need for taking derivative.)
   .

2. (Weighted linear regression) (10 pts) In class when discussing linear regression, we assume that the Gaussian noise is iid (identically independently distributed). In practice, we may have some extra information regarding the fidelity of each data point. For example, we may know that some examples have higher noise variance than others. To model this, we can model the noise variable $\epsilon_i, \epsilon_2, \cdots \epsilon_n$ as distinct Gassians, i.e., $\epsilon_i \sim N(0, \sigma_i^2)$ with known variance σ_i^2. How will this influence our linear regression model? Let's work it out.

   (a) (3pts) Write down the log likelihood function of $\mathbf{w}$ under this new modeling assumption.

   (b) (1pts) Show that maximizing the log likelihood is equivalent to minimizing a **weighted square loss function** $J(\mathbf{W}) = \sum_{i=1}^{n} a_i(\mathbf{w}^T\mathbf{x}_i - y_i)^2$, and express each a_i in terms of σ_i.

   (c) (3 pts) Derive a batch gradient descent update rule for optimizing this objective.

   (d) (3 pts) Derive a closed form solution to this optimization problem. Hint: begin by rewrite the objective into matrix form using a diagonal matrix A with $A(i, i) = a_i$.

3. (Decision theory) (10 pts) For this problem, we will work through an scenario where using the Maximum A-Posteriori decision rule as described in class is not appropriate. Consider a spam filter that gives a probabilistic prediction for each email being a spam. Critically, there is a cost associated with filtering out non-spam emails as well as letting spam email through. But the cost is not symmetric. The following table specifies the costs.

predicted label $\hat{y}$	true label y	
	non-spam	spam
non-spam	0	1
spam	10	0

Table 1: A mis-classification cost matrix for the spam filter problem.

As you can see, if the prediction is correct, there is no cost. But if you mis-classify a non-spam email as a spam, there is a significantly higher cost (10) than the other way around (1).

Here we will go through some questions to help you figure out how to use the probability to make filtering decisions.

(a) (2 pt) For a given email $\mathbf{x}$, the spam filter predicts it being a spam with $p = 0.8$, what is the expected cost of classifying it as *spam*?

(b) (2 pt) We want to minimize the expected cost, how should we classify this particular email?

(c) (3pts) As you can see from parts (a) and (b), using the MAP decision rule (aka comparing p to 0.5) to make prediction for this email is not appropriate and leads to higher expected mis-classification cost. Please devise a decision rule that compares p, the predicted probability of being spam, to a new threshold θ such that we can minimize the expected misclassification cost based on the costs specified in Table 1.

(d) (3pts) Can you provide a cost table for which the decision rule minimizing the expected mis-classification cost would use a $\theta = 1/5$? Please provide the proof that your cost matrix gives $\theta = 1/5$.

4. (8 pts) (Maximum A-Posteriori 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-Posteriori (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) (4 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) (4 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 the prior as well as the two posterior distributions (you can use any software for this). 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?

5. (Perceptron) (4 pts) Assume a data set consists only of a single data point $\{(x, +1)\}$. How many times would the Perceptron algorithm mis-classify this point $\mathbf{x}$ before convergence? What if the initial weight vector $\mathbf{w}_0$ was initialized randomly and not as the all-zero vector?

   (a) (1 pts) Case 1: $\mathbf{w}_0 = 0$.

   (b) (3 pts) Case 2: $\mathbf{w}_0! = 0$ (please derive the solution as a function of $\mathbf{w}_0$ and $\mathbf{x}$.):

6. (Bonus: 8 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:

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

   where y_{ik} is an indicator variable taking value 1 if $y_i = k$. Please compute the log-likelihood function and the gradient of the log-likelihood function w.r.t the weight vector $\mathbf{w}_c$ of class c. [1]

[1] Here are a few tips to help you work through the math. First, the Logistic regression lecture slide actually provides the solution to this problem. But you will need to fill in the missing derivation. Second, for a particular example $\mathbf{x}_i$, the denominator in the softmax function $\sum_j \exp(\mathbf{w}_j^T \mathbf{x}_i)$ is the same for all k. So denoting it as z_i can make it simpler to work through the derivation. But be sure to remember that z_i is a function of all $\mathbf{w}_k$'s)