

AI534 — Written Homework Assignment 1 (30 pts + 6 bonus pts)

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. [3pt]

Given a set of IID observed samples $x_1, \dots, x_n \sim \text{uniform}(0, \theta)$, we wish to estimate the parameter θ .

- (1 pt) Write down the likelihood function of θ .
- (2 pts) Derive the maximum likelihood estimation for θ , which is the value for θ that maximizes the function of part (a). (Hint: The likelihood function is a monotonic function. So the maximizing solution is at the extreme— there is no need to take derivative for this case.)

2. Weighted linear regression. [10pt]

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 $\epsilon_i, \epsilon_2, \dots, \epsilon_n$ as distinct Gaussian's, 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.

- (3pts) Write down the log likelihood function of \mathbf{w} under this new modeling assumption.
- (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 .
- (3 pts) Take the gradient of the loss function $J(\mathbf{w})$ and provide the batch gradient descent update rule for optimizing \mathbf{w} .
- (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: working with expectations. [6pt]

In this problem, you will analyze a scenario where the Maximum A-Posteriori (MAP) decision rule, which you learned in class, is not appropriate. Instead, you'll explore how to make decisions based on minimizing expected costs.

Consider a spam filter that predicts whether an email is spam, using probabilistic predictions. For this filter, there are costs associated with making errors (misclassifying emails), but these costs are not symmetric. Misclassifying a non-spam email as spam (i.e., filtering out an important email) is more costly than misclassifying a spam email as non-spam.

The following table shows the cost of each possible outcome:

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.

- If the filter's prediction is correct, there is no cost.
- If a non-spam email is classified as spam, there is a cost of 10.
- If a spam email is classified as non-spam, there is a cost of 1.

Here we will go through some questions to help you figure out how to use the probability and misclassification costs to make predictions.

- (a) (2 pt) You received an email for which the spam filter predicts that it is a spam with $p = 0.8$. We want to make the decision that minimizes *the expected cost*.

Question: Should you classify this particular email as spam or non-spam? [Hint: Compare the expected cost of classifying the email as spam versus non-spam. Choose the classification that results in the lower expected cost.]

- (b) (2 pts) The MAP decision rule would classify an email as spam if $p > 0.5$, but this rule does not minimize expected cost in this case. We need a new rule that compares p to a different threshold θ . The value of θ should be chosen to minimize the expected cost based on the costs in the table.

Question: What is the value of θ that works for the costs specified in Table 1? [Hint: To find the threshold θ , set up the decision rule by comparing the expected cost of each decision, as you did in (a), then Solve for p in terms of the costs.]

- (c) (2pts) Now, imagine that the optimal decision rule would use $\theta = 1/5$ as the threshold for classifying an email as spam. **Question:** Can you provide a new cost table where this would be the case? [Hint: Use the relationship between the costs and θ that you derived in part (b). Based on this relationship, adjust the misclassification costs in the table to achieve $\theta = 1/5$.]

4. **Maximum A-Posteriori Estimation.** [8pt] Suppose we observe the values of n IID random variables X_1, \dots, 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, \dots, X_n can be viewed as generated by:

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

In this setting, Maximum A-Posteriori (MAP) estimation is a way to estimate θ by finding the value that maximizes the posterior probability, given both its prior distribution and the observed data. The MAP estimation of θ is given by:

$$\hat{\theta}_{MAP} = \underset{\hat{\theta}}{\operatorname{argmax}} P(\theta = \hat{\theta} | X_1, \dots, X_n)$$

By applying Bayes' theorem, this becomes:

$$\hat{\theta}_{MAP} = \underset{\hat{\theta}}{\operatorname{argmax}} P(X_1, \dots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta}) = \underset{\hat{\theta}}{\operatorname{argmax}} L(\hat{\theta}) p(\hat{\theta})$$

where $L(\hat{\theta})$ is the likelihood function of the data given θ , and $p(\hat{\theta})$ is the prior distribution over θ .

Now consider using a beta distribution as the prior: $\theta \sim \text{Beta}(\alpha, \beta)$, 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.

- (a) (3 pts) Derive the posterior distribution $p(\hat{\theta}|X_1, \dots, X_n, \alpha, \beta)$. Compare the form of the posterior distribution with that of the beta distribution, you will see the posterior is also a beta distribution. What the updated α and β parameters for the posterior?
 - (b) (2 pts) Suppose we use $\text{Beta}(2, 2)$ as the prior, what Beta distribution do we get for the posterior 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? (you don't need to write out the distributions, simply provide the α and β distribution would suffice).
 - (c) (1pt) Plot the pdf function of the prior $\text{Beta}(2, 2)$ and the two posterior distributions. You can use any software (e.g., R, Python, Matlab) for this plot.
 - (d) (2pt) Assume that $\theta = 0.4$ is the true probability, as we observe more and more coin tosses from this coin, how would the shape of the posterior change as more data is observed? Will the MAP estimate converge toward the true value?
5. **Perceptron.** [3pt] 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? Derive the number of times as a function of \mathbf{w}_0 and \mathbf{x} .
- (a) (1 pts) Case 1: $\mathbf{w}_0 = 0$.
 - (b) (2 pts) Case 2: $\mathbf{w}_0 \neq 0$:
6. **Bonus: MLE for multi-class logistic regression.** [6 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 $y_i = k$.

- (a) (1 pts) Provide the log-likelihood function.
- (b) (5 pts) Derive the gradient of the log-likelihood function w.r.t the weight vector \mathbf{w}_c of class c . [Hint: the solution to this problem is provided in the Logistic regression lecture slide. You just need to fill in the missing derivation. Note that for any 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 —denoting it as z_i makes it simpler to work through the derivation, but be sure to remember that z_i is a function of all \mathbf{w}_k 's.]