Homework 2

Due : October 17th at 11 : 59 pm

Deliverables. Submit a single PDF of your write-up to Gradescope *HW2 Written Questions*.

Start each problem on a new page.

Honor Code

Write and sign the following statement:

> "I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

1 Sum of Residuals

Consider a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ with $x_n \in \mathbb{R}^{D+1}$ and $y_n \in \mathbb{R}$. Here we assume that X already contains a column of 1s. We model the relationship between x and y using linear regression with parameters $w \in \mathbb{R}^{D+1}$:

$$\hat{y} = x^\top w.$$

We have computed the solution w^* to the least squares regression problem:

$$w^* = \arg\min_w \frac{1}{2} \sum_{n=1}^N \left(y_n - x_n^\top w\right)^2.$$

The residual vector is then defined as:

$$r = y - Xw^*$$

(a) (*2 pts*) Use the normal equation to show that: $X^\top r = 0$.

(b) (*2 pts*) Use the presence of a bias term (encoded in the column of 1s in X) to show that the sum of the residuals is 0.

(c) (*2 pts*) Suppose we add a point (x_j, y_j) to our dataset with a very large $|y_j|$ but x_j is close to the mean of x. Predict qualitatively how w will change and explain why least squares is sensitive to such a point. How can you modify the loss to make the model more robust to outliers?

2 Weighted Linear Regression

Consider a dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ with $x_n \in \mathbb{R}^{D+1}$ and $y_n \in \mathbb{R}$. Here we assume that X already contains a column of 1s. We model the relationship between x and y using linear regression with parameters $w \in \mathbb{R}^{D+1}$:

$$\hat{y} = x^\top w.$$

Suppose we have nonnegative weights $\{\alpha_n\}_{n=1}^N \geq 0$ for each observation and our goal is to minimize the weighted least squares objective:

$$L(w) \;=\; \sum_{i=1}^{n} \alpha_i (y_i - x_i^\top w)^2.$$

In this problem, we will derive the w^* that minimizes this loss.

(a) (*2 pts*) Write the weighted least-squares loss function $L(w)$ in matrix form using the diagonal matrix:

$$A = \text{diag}(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{N \times N}.$$

Your answer, should not contain any summations.

(b) (*2 pts*) Compute the gradient of the loss and express your answer in matrix form (no summations). You can do this by working with the summation form and taking partial derivatives or by using the following gradient identities:

$$\nabla_w \left(u^\top w \right) = u$$
$$\nabla_w \left(w^\top A w \right) = (A + A^\top) w$$

(c) (*3 pts*) Set the gradient equal to zero, and solve for the closed-form solution for w^* that minimizes the weighted least squares objective. in terms of X, y, and A, assuming $X^\top A X$ is invertible.

(d) (*3 pts*) Under what condition is $X^\top A X$ guaranteed to be invertible? Can we derive a simpler condition when $\alpha_i > 0$?

3 Online K-Means

Consider the sequential (online) version of the K-means algorithm. Online algorithms assume that you receive data points in a *stream* rather than having all of your data already stored.

In online K-means, instead of using the full dataset to update cluster centers at each iteration, we update the nearest prototype μ_k for each data point x_n as it is observed, using the rule:

$$\mu_k \leftarrow \mu_k + \frac{1}{N_k}(x_n - \mu_k),$$

where N_k is the number of points that have been assigned to cluster k so far. Each point is assigned to the cluster with the nearest center at the time it is observed.

(a) (*3 pts*) Explain how this sequential update differs from batch K-means in terms of memory and computation. *Hint: How often is each data point used in regular K-means compared to online K-means?*

(b) (*5 pts*) Show that after each update, the objective function

$$E(\{\mu_k\}) = \sum_{n=1}^{N} \min_k \|x_n - \mu_k\|^2$$

does not increase.

4 MLE vs MAP

In this problem, we explore the difference between the Maximum Likelihood Estimate (MLE) and the Bayesian Maximum A Posteriori (MAP) estimate. Suppose we have a coin with unknown bias Y. If we flip the coin we observe a Bernoulli random variable:

$$X \sim \textbf{Bernoulli}(Y) \;=\; p(X = x \mid Y) = Y^x(1-Y)^{1-x}$$

The coin will land heads ($X = 1$) with probability Y and tails ($X = 0$) with probability $1 - Y$. We flip the coin (independently) N times to collect the dataset $\mathcal{D} = \{X_n\}_{n=1}^N$. We observe the coin lands heads $N_1 = \sum_{n=1}^N X_n$ times and tails $N_0 = N - N_1$ times.

For this problem, we will also introduce a prior on the probability Y that the coin lands heads. The Beta distribution is a common prior for probabilities. The Beta distribution has the density:

$$p(y \mid \alpha, \beta) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad y \in (0, 1),$$

where $\alpha > 0$ and $\beta > 0$ are parameters of the distribution. The function $B(\alpha, \beta)$ is the normalization function

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1} dy$$

of the Beta distribution called the Beta function. As a prior we will need to pick values for α and β. These should be based on our beliefs about the coin. Here we will use $\alpha = 3$ and $\beta = 3$:

(a) (*2 pts*) Write the likelihood $p(\mathcal{D}|Y)$ of the data given Y.

(b) (*2 pts*) Compute the derivative of the log-likelihood function with respect to Y and solve for the MLE estimate for Y.

(c) (*6 pts*) Write the posterior distribution $p(Y|\mathcal{D})$. Express this posterior in the form of a Beta distribution. What are the posterior values of α and β based on the prior values of α and β as well as N_0 and N_1?

(d) (*4 pts*) Compute the derivative of the log-posterior and solve for the MAP Estimate.

(e) (*2 pts*) In a 1 sentence, what is the difference between the MLE and MAP estimates? Note: simply putting the your solutions to the previous parts is not sufficient.

(f) (*5 pts*) How would you increase or decrease the regularization by varying α and β? *Hint: Regularization was covered in the bias variance trade off lectures: more regularization results in a higher bias and lower variance.*