In the name of GOD.

Sharif University of Technology

Machine Learning

Spring 2025

Hamid R. Rabiee, Zahra Dehghanian

Homework 1	Regression, Bias and Variance, ML/MAP	Deadline: 1403/12/10

1. [18] In the real world, we are often faced with problems where the data follows the regression function $h(x)$. That is, for any input value x, the corresponding label is $y = h(x) + \epsilon$, where ϵ represents noise. Using the given data, we create M predictive models $y_1(x), y_2(x), \ldots, y_M(x)$. If $H_M(x)$ is defined as:

$$H_M(x) = \frac{1}{M} \sum_{i=1}^{M} y_i(x),$$

prove that:

$$\mathbb{E}_x \left[(H_M(x) - h(x))^2 \right] \leq \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_x \left[(y_i(x) - h(x))^2 \right].$$

hint:Cauchy-Schwarz!

2. [20] Consider a standard linear regression setting where there exists a true parameter vector $\theta^* \in \mathbb{R}^{d \times 1}$ that generates outputs $y_i \in \mathbb{R}$ from inputs $x_i \in \mathbb{R}^{d \times 1}$ as follows:

$$y_i = {\theta^*}^T x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Suppose we have fixed the inputs $x_1, \ldots, x_n$. We construct the matrix X by placing x_i in the i-th row ($X \in \mathbb{R}^{n \times d}$). The training error for a given parameter θ can be written in matrix form as $\hat{R}(\theta) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\theta\|_2^2$. We define the statistical error of parameter θ, denoted as $R(\theta)$, as the expectation of the training error over the randomness in $\mathbf{y}$, that is:

$$R(\theta) = \mathbb{E}_{\mathbf{y}}[\hat{R}(\theta)].$$

Since the design matrix $\mathbf{X}$ is fixed, the only source of randomness comes from ϵ. Let R^* be the statistical error corresponding to the true parameter θ^*, which generates the outputs $\mathbf{y}$. Imagine that the matrix $\mathbf{X}$ is full rank, and the OLS estimator is written as $\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Prove the following:

(a) Show that

$$R(\theta) - R^* = \|\theta - \theta^*\|_{\hat{\Sigma}}^2,$$

where $\|\theta - \theta^*\|_{\hat{\Sigma}}^2 = (\hat{\theta} - \theta^*)^T \hat{\Sigma} (\hat{\theta} - \theta^*)$ and $\hat{\Sigma} = \frac{\mathbf{X}^T\mathbf{X}}{n}$ is the empirical covariance matrix.

(b) Prove that the OLS estimator is unbiased. Additionally, show that the variance of the OLS estimator $\hat{\theta}$ is:

$$\text{var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta^*)(\hat{\theta} - \theta^*)^T] = \frac{\sigma^2}{n}\hat{\Sigma}^{-1}.$$

3. [12] Consider a linear regression model where the cost function is given by:

$$J = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \mathbf{w}^T\mathbf{w}$$

where:

- $\mathbf{y}$ is the target vector,
- $\mathbf{X}$ is the feature matrix,
- $\mathbf{w}$ is the weight vector,
- λ is the regularization parameter,
- $\|\cdot\|_2$ denotes the L_2 norm.

(a) Explain what $\lambda\mathbf{w}^T\mathbf{w}$ is and why we add it to the loss function. How does it work?

(b) Explain the importance of selecting an appropriate value for the regularization parameter λ. What happens if we choose λ very large (i.e., $\lambda \to \infty$) or very small (i.e., $\lambda \to 0$)?

(c) Consider the general form of regularization:

$$J = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda r(\mathbf{w}),$$

where $r : \mathbb{R}^d \to \mathbb{R}$ is the regularization function. Suppose for two weight vectors $\mathbf{w}_1 = [1, 0, 6]^T$ and $\mathbf{w}_2 = [-2, 3, 3]^T$, the squared error term $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ is equal for both. Explain which weight vector ($\mathbf{w}_1$ or $\mathbf{w}_2$) the model will prefer if we replace $r(\mathbf{w})$ with each of the following functions:

- $r(\mathbf{w}) = 1$,
- $r(\mathbf{w}) = \sum_{i=1}^{d} |w_i|$ (L_1 regularization),
- $r(\mathbf{w}) = \sum_{i=1}^{d} |w_i|^2$ (L_2 regularization),
- $r(\mathbf{w}) = \max(|w_1|, |w_2|, \ldots, |w_d|)$ (L_∞ regularization).

(d) Suppose we are using **Stochastic Gradient Descent (SGD)** to optimize the cost function:

$$J = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \mathbf{w}^T\mathbf{w}.$$

Derive the update rule for the weight w_i, using learning rate η and data point $(\mathbf{x}^{(j)}, y^{(j)})$ in SGD.

4. [35] Answer the following quesitons.

(a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from $\text{Ber}(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for θ?

(b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

(c) Assume Θ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1$, $\pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

(d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0, 1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0, 1)$ instead of our discrete one. We assign $\Theta \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density

$$\pi_\Theta(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } \theta \in (0, 1).$$

Recall the mode of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is

$$\frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

(the mode is the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_\Theta(\theta \mid x)$ is proportional to $\text{Beta}(k+\alpha, n-k+\beta)$ distribution, and find the MAP estimator.

(f) Recall that $\text{Beta}(1, 1) \equiv \text{Unif}(0, 1)$ (pretend we saw $1 - 1$ heads and $1 - 1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

(g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter p. Interpret α, β as to how they affect our estimate. This is a really special property: if the prior distribution multiplied by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

(h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when n is small, or n is large?

(i) Which do you think is "better", MLE or MAP?

5. [15] Complete the attached notebook. You are permitted to use chatbots or other resources for assistance, but you must ensure that you fully understand the code and implementations. During the online sessions, you may be asked to explain specific functions, code lines, and your overall approach. Be prepared to demonstrate your understanding in detail.