Machine Learning 2024 - Sheet 2.1

Isabel Valera

Notation. The input feature vector of the i-th sample, i.e., $\mathbf{x}_i \in \mathbb{R}^D$, can be used to construct feature matrix for N samples represented as $\mathbf{X} \in \mathbb{R}^{N \times D}$. The input vector can be represented by a basis function $\Phi(\mathbf{x}_i) \in \mathbb{R}^M$. The feature matrix for N samples is then represented as $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$. The target vector of the i-th sample, i.e., $y_i \in \mathbb{R}$, can be used to construct column vector for N samples represented as $\mathbf{Y} \in \mathbb{R}^N$.

Exercise 1: Orthogonal Projection ⚡⚡

i) Show that the matrix $\mathbf{A}$ in Equation 1 takes any vector $\mathbf{v}$ and projects it (i.e. $\mathbf{Av}$) onto the space spanned by the columns of $\mathbf{\Phi}$.

$$\mathbf{A} = \mathbf{\Phi} \left(\mathbf{\Phi}^\top \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^\top \tag{1}$$

Note: Definition of the column space of matrix $\mathbf{A}_{m \times n}$:

$$\text{Col}(\mathbf{A}) = \{ \mathbf{b} \in \mathbb{R}^m \,|\, \mathbf{Ax} = \mathbf{b} \} \tag{2}$$

ii) Use this result to show that the least-squares predictions $\hat{\mathbf{Y}} = \mathbf{\Phi} \mathbf{w}_n$, where $\mathbf{w}_n$ is the empirical solution given in Equation 3, correspond to an **orthogonal projection** of the vector $\mathbf{Y} \in \mathbb{R}^N$ onto the manifold $\mathcal{S}$ as shown in Figure 1.

Hint: Manifold $\mathcal{S}$ is spanned by the column vectors of $\mathbf{\Phi}$.

$$\mathbf{w}_n = \left(\mathbf{\Phi}^\top \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^\top \mathbf{Y} \tag{3}$$

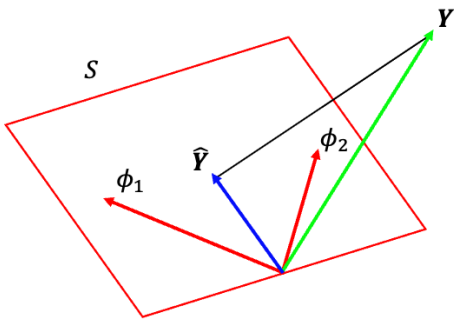

Figure 1: Geometric interpretation of least squares solution in an n-dimensional space. Least square regression function Predictions from least squares regression $\hat{\mathbf{Y}} = \mathbf{\Phi} \mathbf{w}_n$ correspond to projection of $\mathbf{Y}$ on to subspace $\mathcal{S}$.

Exercise 2: Least squares regression ⚡

Recall the general form for the least squares regression problem:

$$\mathbf{w}_n = \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2$$

Given a training dataset (i.e, (x_i, y_i) pairs) $\{(10, 22), (1, 2.5), (2, 3)\}$.

i) Compute the optimal weights and the predictions for the training data.

ii) Suppose now that $w_0 = 0$ and $w_1 = 2$. Evaluate the gradient of the error w.r.t. w_1 using the training data.

Solutions:

i)

$$\mathbf{w}_n^* = \begin{bmatrix} -83/146 \\ 328/146 \end{bmatrix}$$

ii)

$$\frac{\partial L}{\partial w_1} = -12.333$$

Exercise 3: Weighted/Repeated Samples ⚡⚡

Consider a data set in which each data point y_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares loss function is as in Equation 5. We will assume that target variable is given by a deterministic function with additive Gaussian noise:

$$y_n = \mathbf{w}^\top \Phi(\mathbf{x}_n) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n \left(y_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\right)^2 \tag{5}$$

i) Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function.

ii) Show the relationship between r_n and σ^2.

Hint: Maximizing the loglikelihood in Equation 6 w.r.t. $\mathbf{w}$ is equivalent to minimizing the weighted sum-of-squares error function. Compare those terms.

$$\ln p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}(y_n|\mathbf{w}^\top \Phi(\mathbf{x}_n), \sigma^2) \tag{6}$$

Solutions:

i) $\mathbf{w}^* = \left(\mathbf{\Phi'}^\top \mathbf{\Phi'}\right)^{-1} \mathbf{\Phi'}^\top \mathbf{Y'}$ where our targets and features are as following

$$\mathbf{Y'} = [\sqrt{r_1}y_1, \sqrt{r_2}y_2, ..., \sqrt{r_N}y_N]^\top \quad \mathbf{\Phi'} = \begin{bmatrix} \sqrt{r_1}\Phi(\mathbf{x}_1) \\ \vdots \\ \sqrt{r_N}\Phi(\mathbf{x}_N) \end{bmatrix}_{N\times M}$$

ii) The weighting factor r_n can be thought as re-scaling the variance of additive Gaussian noise ($\sigma^2 \to \frac{1}{r_n}\sigma^2$) assumed in Equation 4.

Exercise 4: Independent noise and weight regularization ⚡⚡⚡

First consider a dataset where $D = \{\mathbf{x}_n, y_n\}_{n=1}^N$ $\mathbf{x}_n \in \mathbb{R}^d, y_n \in \mathbb{R}$. Now assume that there is also a noisy version of the dataset $\widetilde{D} = \{\tilde{\mathbf{x}_n}, \tilde{y}_n\}_{n=1}^N$ where

$$\tilde{\mathbf{x}}_n = \mathbf{x}_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \quad \texttt{and} \quad \tilde{y}_n = y_n \quad \forall n \in \{1, ..., N\} \tag{7}$$

where $\epsilon_n \in \mathbb{R}^d$. Consider the linear model given in equation 8 together with squared loss function given in 9.

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i \tag{8}$$

$$L(y_n, \mathbf{x}_n, \mathbf{w}) = \left(y_n - f(\mathbf{x}_n, \mathbf{w})\right)^2 \tag{9}$$

i) Show that

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}_\epsilon\Big[\mathbb{E}_{\widetilde{D}}\big[L(X, Y, \mathbf{w})\big]\Big] = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}_D\big[L(X, Y, \mathbf{w})\big] + \lambda\Omega(\mathbf{w}) \tag{10}$$

where X, Y are random variables, the LHS term corresponds to **empirical risk minimization (ERM)** and the RHS term denotes **regularized ERM**.

Note: $\mathbb{E}[\epsilon_i] = 0, \quad \mathbb{E}[\epsilon_i \epsilon_j] = \mathbb{1}_{i=j}\sigma^2 \quad \forall i, j.$

In this part of the question, assume that true data generation is given by

$$y_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{11}$$

and we can sample a dataset $D = \{\mathbf{x}_n, y_n\}_{n=1}^N$. We estimate weights by using regularized ERM objective by using the linear model given in Equation 8.

ii) How does **omitting** the coefficient ($\lambda = 0$) of weight regularization term affect the bias and the variance of the predictions made by the resulting regressor?

iii) How does **increasing** the coefficient ($\lambda \gg 0$) of weight regularization term affect the bias and the variance of the predictions made by the resulting regressor?

iv) How does **a negative valued** ($\lambda \ll 0$) weight regularization parameter affect the bias and the variance of the predictions made by the resulting regressor?

v) Assume that we are not limited to fixed size training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, instead we can be in infinite data regime $N \to \infty$. Please comment on the advantage or disadvantage of using a weight regularization term in this case (hint: consider the motivation of using a regularizer and its effect on the variance of the models).

Solutions:

ii) When $\lambda = 0$, our objective becomes empirical risk minimization which corresponds finding least squares estimation of $\mathbf{w}$.
iii) When we introduce $\lambda \gg 0$, we regularize the norm of the weights in our objective. This will increase the bias of the weights, which also increases the bias of the predictions. However, variance of the predictions will decrease. So, compared to previous case, we will observe an increase in bias and decrease in variance.
iv) This has no effect of regularization and will create infinite weights.
v) If you can access datapoints when $N \to \infty$ (which is never the case in reality), you do not have to worry about using a regularizer because you can directly train a model that can already generalize to all datapoints.

Exercise 5: Linear Models and Basis functions ⚡⚡⚡

Consider a linear basis function regression model for a multivariate target variable $\boldsymbol{y}$ having a Gaussian distribution as given in Equation 12, where $f(\mathbf{x}, \mathbf{W}) = \mathbf{W}^\top \Phi(\mathbf{x})$ together with a training data set comprising transformed input features $\Phi(\mathbf{x}_n)$ and corresponding target vectors $\mathbf{y}_n$, with $n = 1, ..., N$. We can write the probability density function for a random target variable as

$$p(\mathbf{y} \,|\, \mathbf{x}, \mathbf{W}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{y} \,|\, f(\mathbf{x}, \mathbf{W}), \boldsymbol{\Sigma}) \tag{12}$$

i) Show that the maximum likelihood solution $\mathbf{W}_{ML}$ for the parameter matrix $\mathbf{W}$ has the property that each column is given by an expression as in Equation 3. Note that this is independent of the covariance matrix $\boldsymbol{\Sigma}$.

ii) Show that the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{y}_n - \mathbf{W}_{ML}^\top \Phi(\mathbf{x}_n)\right) \left(\mathbf{y}_n - \mathbf{W}_{ML}^\top \Phi(\mathbf{x}_n)\right)^\top$$

iii) Write down a set of basis functions $\{\phi_k(x)\}_k$ that allows you to represent the months of the year $x \in \{1, 2, 3, ..., 12\}$ such that the Euclidean distance between the resulting embedding for any two consecutive months (e.g., March and February, January and December, and so on) is the same.

iv) Assume that you have two variables x_1, x_2 with a scalar target variable y and we want to make a quadratic fit using them, namely,

$$\hat{y} = f\left(x_1, x_2, w_0, w_1, w_2, w_3, w_5\right) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 \left(x_1\right)^2 + w_5 \left(x_2\right)^2 \tag{13}$$

Propose an approach to reformulate our regressor f so that it stays linear in the $\mathbf{w}$, i.e. $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \Phi(\mathbf{x})$ where $\mathbf{w} = [w_0, w_1, w_2, w_3, w_5]^\top$.

Solutions:

iii) We can use trigonometric functions to map each x to evenly distributed points on a circle.

$$\phi(x) = \left\{ \underbrace{\cos\left(x * \frac{2\pi}{12}\right)}_{\phi_1(x)}, \underbrace{\sin\left(x * \frac{2\pi}{12}\right)}_{\phi_2(x)} \right\}$$

iv) We can come up with the following change of basis where $\mathbf{x} = [x_1, x_2]^\top$

$$\phi_0(\mathbf{x}) = 0, \quad \phi_1(\mathbf{x}) = x_1, \quad \phi_2(\mathbf{x}) = x_2, \quad \phi_3(\mathbf{x}) = x_1 x_2, \quad \phi_4(\mathbf{x}) = (x_1)^2, \quad \phi_5(\mathbf{x}) = (x_2)^2$$

Hence, we can write our new $\mathbf{x}' = [0, x_1, x_2, x_1 x_2, x^2{}_1, x^2{}_2]^\top$ with corresponding $\mathbf{w} = [w_0, w_1, w_2, w_3, w_5]^\top$. Now, we can perform linear regression, i.e. $f(\mathbf{x}', \mathbf{w}) = \mathbf{w}^\top \mathbf{x}'$ and optimise our weights by using least squares.

Exercise 6: Gauss-Markov theorem ⚡⚡⚡

Assume that we have datapoints, each of them obeys to linear model

$$y = \mathbf{w}^\top \Phi(\mathbf{x}) + \epsilon \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

where $\Phi(\mathbf{x}) \in \mathbb{R}^M, y \in \mathbb{R}$ and ϵ at different points are uncorrelated.

i) Let $\theta = \boldsymbol{\alpha}^\top \mathbf{w}$ be a linear combination of the parameters $\boldsymbol{\alpha}, \mathbf{w}$. Prove the Gauss-Markov theorem: the least squares estimates of θ, i.e. $\theta^* = \boldsymbol{\alpha}^\top \mathbf{w}^*$, has a variance no bigger than that of any other linear unbiased estimate of θ, i.e. $\tilde{\theta}$.

ii) Move to multivariate case, you can assume that your data follow the model in Equation 15. Show that if $\boldsymbol{\Sigma}^*$ is the covariance matrix of the least squares estimate of $\mathbf{w}$ and $\tilde{\boldsymbol{\Sigma}}$ is the covariance matrix of any other linear unbiased estimate, then $\boldsymbol{\Sigma}^* \preceq \tilde{\boldsymbol{\Sigma}}$.

$$\mathbf{y} = \mathbf{W}^\top \Phi(\mathbf{x}_n) + \epsilon \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{15}$$

where $\Phi(\mathbf{x}) \in \mathbb{R}^{M \times 1}, \mathbf{y} \in \mathbb{R}^{K \times 1}, \mathbf{W} \in \mathbb{R}^{M \times K}$.

Note: The matrix inequality $\boldsymbol{B} \preceq \boldsymbol{A}$ holds if $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite.

Exercise 7: Ridge regression ⚡⚡

i) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\mathbf{w} \sim \mathcal{N}(0, \tau^2 \boldsymbol{I})$ and Gaussian sampling model $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \boldsymbol{I})$.

ii) Find the relationship between the regularization parameter λ in the ridge formula 16, and the variances τ^2 and σ^2.

$$L(\mathbf{X}, \mathbf{Y}, \mathbf{w}) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{D} x_{ij} w_j \right)^2 + \lambda \sum_{j=1}^{D} w_j^2 \tag{16}$$

Solutions:

ii) $\lambda = \frac{\sigma^2}{\tau^2}$.

Artificial data3 Assume that we have datapoints that obeys to linear model

$$\mathbf{Y} = \mathbf{X}\mathbf{w} + \epsilon \tag{17}$$

where $\mathbf{X} \in \mathbb{R}^{N \times D}$ (each row is a feature), $\mathbf{w} \in \mathbb{R}^{D \times 1}$, $\mathbf{Y} \in \mathbb{R}^{N \times 1}$ (each row is a target). Show that the ridge regression estimates of weights, i.e. $\mathbf{w}_{RR}$, can be obtained by ordinary least squares regression on an augmented data set where the dataset matrix $\mathbf{X}_{N \times D}$ is augmented with additional matrix $\mathbf{A}_{D \times D} = \sqrt{\lambda} \mathbf{I}_{D \times D}$, and the target matrix $\mathbf{Y}$ is augmented with matrix $\mathbf{0}_{D \times 1}$ as in Equation 18.

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{A} \end{bmatrix}_{(N+D) \times D} \quad \widetilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{0} \end{bmatrix}_{(N+D) \times 1} \tag{18}$$

Note: By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients towards zero. This is related to the idea of *hints* due to Abu-Mostafa (1995), where model constraints are implemented by adding artificial data examples that satisfy them.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*. Springer series in statistics New York, 2001.