Section 02: Maximum Likelihood Estimation, Linear Algebra, Linear Regression, and Bias-Variance Tradeoff

In this section, we explore maximum likelihood estimation with more examples of noise densities; we review some basics about subspaces in linear algebra; we explore a general version of linear regression, going over the proof in two different formats (matrix and coordinate); and we study bias-variance trade-off.

1. Maximum Likelihood Estimation

In this section, we formulate maximum likelihood estimation for different noise densities as different minimization problems. Specifically, we'll see how each noise distribution corresponds to a specific objective function.

We consider the linear measurement model (parameterized by w), $y_i = x_i^\top w + v_i$ for $i = 1, 2, \ldots, m$. The noise v_i for different measurements (x_i, y_i) are all independent and identically distributed. Under our assumption of a linear model, $v_i = y_i - x_i^\top w$. Note Per the principle of maximum likelihood estimation, we seek to maximize

$$\log p_w((x_1, y_1), \cdots, (x_m, y_m)) = \log \prod_{i=1}^{m} p(y_i - x_i^\top w).$$

(a) Show that when the noise measurements follow a Gaussian distribution ($v_i \sim \mathcal{N}(0, \sigma^2)$), the maximum likelihood estimate of w is the solution to $\min_w \|Xw - Y\|_2^2$. Here each row in X corresponds to a x_i, and each row in Y to y_i.

(b) When the noise measurements follow a Laplacian distribution ($p(z) = (1/2a)\exp(-|z|/a)$), what is the maximum likelihood estimate of x? Express your answer as the solution to an optimization problem such as in the previous part.

(c) When the noise measurements follow a uniform distribution ($p(z) = (1/2a)$ on $[-a, a]$), what is the maximum likelihood estimate of w? Express your answer as a condition to be satisfied by some function of w.

2. Linear Algebra Review

Let $X \in \mathbb{R}^{m\times n}$. X may not have full rank. We explore properties about the four fundamental subspaces of X.

2.1. Summation form v.s. Matrix form

(a) Let $w \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Let x_i denotes each row in X and y_i in Y. Show $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$

(b) Let $L(w) = \|Xw - Y\|_2^2$. What is $\nabla_w L(w)$? (Hint: You can use either summation or matrix form from first sub-problem).

2.2. Subspaces of X

What is the rowspace, columnspace, nullspace, and rank of $X = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6. \end{pmatrix}$.

2.3. Connections between subspaces of X

Check the following facts.

(a) The rowspace of X is the columnspace of $X^\top$, and vice versa.

(b) The nullspace of X and the rowspace of X are orthogonal complements. This can be written in shorthand as Null$(X) =$ Range$(X^\top)^\perp$. This is further equivalent to saying Range$(X^\top) =$ Null$(X)^\perp$.

(c) The nullspace of $X^\top$ is orthogonal to the columnspace of X. This can be written in shorthand as Null$(X^\top) =$ Range$(X)^\perp$.

2.4. Linear algebra facts for linear regression

We saw in lecture on Linear Regression that the closed form expression for linear regression without an offset involves the term $(X^\top X)^{-1}$.

(a) Is it true that the matrix $X^\top X$ is always symmetric and positive semidefinite?

(b) State and prove the connection between the nullspace of X and the nullspace of $X^\top X$. That is, your statement should look like one of the following: Null$(X) \subseteq$ Null$(X^\top X)$, or Null$(X) \supseteq$ Null$(X^\top X)$ or Null$(X) =$ Null$(X^\top X)$.

(c) Is it true that $X^\top X$ is always invertible?

(d) Based on the above fact about the connection between the nullspaces of X and $X^\top X$ and the expression for linear regression without an offset (that we referred to two problems above), justify the use of "tall skinny" data matrix X as opposed to a "short wide" matrix X.

(e) The columnspace and rowspace of $X^\top X$ are the same, and are equal to the rowspace of X. (Hint: Use the relationship between nullspace and rowspace.)

3. Bias-Variance Trade-off

Consider a simple statistical learning setting, in which we assume that there is some unknown function relating two random variables X and Y (e.g. $Y = 2X$). Let us denote this function by $Y = \eta(X)$; however, we don't know specifically what this function $\eta(\cdot)$ is. Our goal is as follows. Given X, we want to predict Y with the smallest possible error, in expectation. We formalize this notion below.

(a) Find the function η that minimizes the expected squared error $\mathbb{E}[(Y - \eta(X))^2]$. **Hint:** Observe that $\mathbb{E}[(Y - \eta(X))^2] = \mathbb{E}[\mathbb{E}[(Y - \eta(X))^2|X = x]]$ (The "Tower Rule").

(b) While ideally we want η to be what we computed above, in reality, however, we are restricted to our training data and a function class, the best we can do is
$\hat{f}_D = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$, where $D = \{(x_i, y_i)\}$. Here, (x_i, y_i) is a sample from distribution P_{XY}. To account for the prediction error (i.e. quality of our estimator $\hat{f}_D$), we need to calculate

$$\mathbb{E}[\mathbb{E}_D[(Y - \hat{f}_D(x))^2]|X = x]]$$

We can break the expectation into

$$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2|X = x]] + \mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$$

$\mathbb{E}[\mathbb{E}[(Y - \eta(x))^2|X = x]]$ is called **irreducible error —** the error incurred even in ideal situation.

$\mathbb{E}_D[(\eta(x) - \hat{f}_D(x))^2]$ is called **learning error —** the error incurred by the learning setting (e.g. insufficient data, the chosen model class F is not expressive enough etc.)

Express the **learning error** in terms of

- bias — $(\eta(x) - \mathbb{E}_D[\hat{f}_D(x)])$
- and variance — $\mathbb{E}_D[(\mathbb{E}_D[\hat{f}_D(x)] - \hat{f}_D(x))^2]$

and explain why there is a trade-off.

Hint: $\eta(x) = \theta$, $\hat{f}_D(x) = \hat{\theta}$ and $\mathbb{E}[\hat{f}_D(x)] = \theta^*$

4. Generalized Least Squares Regression

We already saw linear regression in class and the ridge regression will be covered in week three. Here we consider a problem that generalizes both of these. As a reminder, in linear regression, we seek a model that captures a linear relationship between input data and output data. The general case we consider imposes additional structure on the model.

Consider an experiment in which you have n data points $x_i \in \mathbb{R}^d$ and corresponding n observations y_i. We wish to come up with a model $\omega \in \mathbb{R}^d$ that satisfies the following properties: first, the error $\sum_{i=1}^n (x_i^\top \omega - y_i)^2$ should be small; second, we don't want small changes in training data resulting in large changes in solution; third, we want to put different weights in controlling the magnitude of different coordinates of ω. We therefore define

$$\widehat{\omega}_{\text{general}} = \arg\min_{\omega} \sum_{i=1}^{n} (y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii} \omega_i^2.$$

Here, D is a diagonal matrix, with positive entries on the diagonal. Observe that when D is the identity matrix, we recover ridge regression, and when $\lambda = 0$, we recover least squares regression. Different weights on D_{ii} cause the magnitudes of ω_i to be controlled differently.

4.1. Closed form in the general case

Deduce the closed form solution for $\widehat{\omega}_{\text{general}}$. You should be comfortable with proofs in the ”coordinate” form as well as the ”matrix” form.

4.2. Special cases: linear regression and ridge regression

(a) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double all the values of y_i?

(b) In the simple least squares case ($\lambda = 0$ above), what happens to the resulting $\widehat{\omega}$ if we double the data matrix $X \in \mathbb{R}^{n \times d}$?

(c) Suppose $D = I$ (that is, it is the identity matrix). That is, this is the *ridge* regression setting. Explain why $\lambda > 0$ ensures a "well-conditioned" setting.