ECE 8803: Online Decision Making in Machine Learning **Fall 2022**

Lecture 2: August 25

Lecturer: Vidya Muthukumar

In this note, we first review the basics of the linear data model and the ordinary-least-squares (OLS) estimator. Then, we introduce the definition of unbiased estimates. Throughout this note, we denote random variables by capital letters (e.g. X, Y, Z) and vectors by boldface (e.g. $\mathbf{x}, \mathbf{y}, \mathbf{z}$).

2.1. Linear Model and OLS estimator

While this class is going to primarily focus on the *online* learning problem, where data arrives in a stream, it is useful to review some basics of the more classical *offline* machine learning paradigm. In this setup, we are given a set of n training examples $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector and $y \in \mathbb{R}$ is the target value. We will consider the case $d < n$ throughout this note.

Linear model: In a linear (regression) model, we assume that y can be expressed as a linear function of $\mathbf{x}$ plus independent noise. In other words, for every $i = 1, \ldots, n$, we have:

$$y_i = \mathbf{w}_*^\top \mathbf{x}_i + \epsilon_i,$$

where $\mathbf{w}_* \in \mathbb{R}^d$ is an unknown parameter vector, and ϵ_i is a random variable that models the noise. Under this model, we are interested in estimating $\mathbf{w}_*$ given access to the training set $\mathcal{D}$. The following is a simple example in the 1-dimensional case, where we would like to model the relationship between x and y with a linear function.

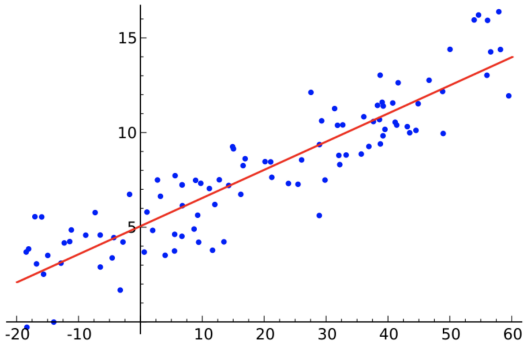

Figure 2.1: Linear regression in 1 dimension.

Squared estimation error: One of the most natural methods to estimate $\mathbf{w}_*$ is to choose the parameter that minimizes the squared estimation error over training data. In other words, we pick

$$\widetilde{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w}) = \sum_{i=1}^{n} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 , \tag{2.1}$$

where $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \dots & \mathbf{x}_n\end{bmatrix}^\top$, and $\mathbf{y} = \begin{bmatrix}y_1 & \dots & y_n\end{bmatrix}^\top$. The estimator $\widetilde{\mathbf{w}}$ is often called the *ordinary-least-squares* (OLS) estimator.

Why is the OLS estimator a natural choice? Here we provide one explanation: *minimizing the squared estimation error is in fact equivalent to maximum likelihood estimation* when $\boldsymbol{\epsilon}$ follows the normal distribution. In more detail: suppose that $\{\epsilon_i\}_{i=1}^n$ are independent and identically distributed (iid) and follow the normal distribution $\mathcal{N}(0, \sigma^2)$. Then, according to the linear model, we have $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ and the y_i's are iid. Thus, the log-likelihood function is given by

$$L(\mathbf{w}) = \ln\left(\prod_{i=1}^{n} P(y_i|\mathbf{w}, \mathbf{x}_i)\right) = -\frac{1}{2\sigma^2}\sum_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 - n\ln\sigma\sqrt{2\pi}.$$

It follows from the above that *maximizing* $L(\mathbf{w})$ is equivalent to *minimizing* the squared estimation error.

An explicit expression for the OLS estimator: We now show that the OLS estimator $\widetilde{\mathbf{w}}$ is unique and can be computed in closed form: this is a special property of the square loss. We note that $f(\mathbf{w})$ is strictly convex[1] in $\mathbf{w}$. Therefore, its minimizer $\widetilde{\mathbf{w}}$ must satisfy the condition $\nabla f(\widetilde{\mathbf{w}}) = \mathbf{0}$. Standard matrix-vector operations then give us

$$2\mathbf{X}^\top \mathbf{X}\widetilde{\mathbf{w}} - 2\mathbf{X}^\top \mathbf{y} = \mathbf{0}.$$

Next, we assume that $\mathbf{X}^\top \mathbf{X}$ is invertible. Then, the OLS estimator $\widetilde{\mathbf{w}}$ can be directly obtained as

$$\widetilde{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

Finally, we note that the OLS estimator is unique. This follows since $\mathbf{X}^\top \mathbf{X}$ is invertible, and so the Hessian matrix $\nabla^2 f(\mathbf{w}) = 2\mathbf{X}\mathbf{X}^\top \succ 0$. This implies that there only exists *one point* where the gradient is zero, and so the OLS solution is unique.

Ridge regression We just computed the unique OLS estimator in closed form for the case when $\mathbf{X}^\top \mathbf{X}$ is invertible. If this is not the case, the OLS estimator need not be unique (and may suffer from poor test performance). In this situation, it makes sense to add extra regularization to the optimization problem to make the solution more "stable". One way of doing this is to add the so-called "ridge regularization" as below:

$$\widetilde{\mathbf{w}}_\lambda = \operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w}) = \sum_{i=1}^{n} \left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2 + \lambda\|\mathbf{w}\|^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2. \tag{2.2}$$

1. We will cover more details on convexity and (strict or strong) convexity a few weeks into the semester.

Following a similar derivation as for the OLS estimator, we now get

$$\widetilde{\mathbf{w}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Note that $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is invertible for any $\lambda > 0$ regardless of the matrix $\mathbf{X}$. This estimator is commonly called the *ridge regression* estimator, and its performance in practice typically depends on the set value of λ.

2.2. Unbiased estimator

Let $\boldsymbol{\beta}$ be a parameter, and $\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$ (typically constructed from random data). Then $\hat{\boldsymbol{\beta}}$ is an unbiased estimation of $\boldsymbol{\beta}$ if

$$\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$

This property of unbiasedness is natural to have in a statistical estimator (although it is also important for the *variance* of the estimator to be low, we do not elaborate on this point in this note).

We now review three natural examples of unbiased estimators that commonly arise in practice.

Example 1 (unbiasedness of the sample mean): Let $\mathcal{D}$ be any distribution, and let $\{X_i\}_{i=1}^n$ be a set of i.i.d. random variables generated from $\mathcal{D}$. Then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimation of μ, since

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n} n\mu = \mu.$$

Recall that we considered this sample mean in the review note on probability and statistics, and showed that it concentrates around its expectation μ via Hoeffding's inequality.

Example 2 (unbiasedness of OLS estimator): In the linear model, the OLS estimator $\widetilde{\mathbf{w}}$ is an unbiased estimator of $\mathbf{w}_*$, if the noise is zero mean. To prove this, note that:

$$\widetilde{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\mathbf{w}_* + \boldsymbol{\epsilon}) = \mathbf{w}_* + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}.$$

Thus, $\mathrm{E}[\tilde{\mathbf{w}}] = \mathbf{w}_*$ as long as $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.

Example 3 (unbiasedness of "importance-weighted" estimator): Suppose we have an example set with n examples: $\mathcal{D} = \{x_1, x_2, \ldots, x_K\}$, and the value of the examples are unknown. Now, we sample J from $\{1, \ldots, K\}$ under a distribution p and observe only x_j. Let $p(i)$ denote the probability that x_i is chosen. We now show that for all $i \in \{1, \ldots, K\}$, the estimator

$$\hat{X}_i = \begin{cases} x_i / p(i), & J = i, \\ 0, & \text{otherwise} \end{cases}$$

is an unbiased estimation of x_i. To prove this, not that i is a random variable, so

$$\mathrm{E}[\hat{X}_i] = p(i) \cdot \frac{x_i}{p(i)} + (1 - p(i)) \cdot 0 = x_i.$$

This type of unbiased estimator will be critically used to construct certain natural multi-armed bandit algorithms later in the course.