Section 02: Solutions

Curated by. Leo Liu (zeyuliu2@cs.washington.edu) and Kyle Zhang (guohaz@cs.washington.edu)

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

Contents

1 Maximum Likelihood Estimation **2**

2 Linear Algebra Review **4**
2.1 Summation form v.s. Matrix form 4
2.2 Subspaces of X 4
2.3 Connections between subspaces of X 4
2.4 Linear algebra facts for linear regression 5

3 Bias-Variance Trade-off **6**

4 Generalized Least Squares Regression **8**
4.1 Closed form in the general case 8
4.2 Special cases: linear regression and ridge regression 9

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.

Solution:

When $v_i \sim \mathcal{N}(0, \sigma^2)$, the density is given by the expression $p(v) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-v^2/2\sigma^2}$. This implies that the MLE of parameter is

$$\begin{aligned}
\hat{w}_{MLE} &= \arg\max_w \log p_w((x_1, y_1), \cdots, (x_m, y_m)) \\
&= \arg\max_w \log \prod_{i=1}^{m} p(y_i - x_i^\top w) \\
&= \arg\max_w \sum_{i=1}^{m} \log p(y_i - x_i^\top w) \quad [\log(ab) = \log a + \log b] \\
&= \arg\max_w \sum_{i=1}^{m} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i^\top w)^2/2\sigma^2} \right] \\
&= \arg\max_w m \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \sum_{i=1}^{m} -\frac{(y_i - x_i^\top w)^2}{2\sigma^2} \\
&= \arg\max_w \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(y_i - x_i^\top w)^2 \quad \text{(constant offset doesn't affect results)} \\
&= \arg\max_w \sum_{i=1}^{m} -(y_i - x_i^\top w)^2 \quad \text{(constant scalar doesn't affect results)} \\
&= \arg\min_w \sum_{i=1}^{m} (y_i - x_i^\top w)^2 = \arg\min_w \|Xw - Y\|_2^2
\end{aligned}$$

Therefore, the maximum likelihood estimate of w is $\arg\min \|Xw - Y\|_2^2$, as claimed.

(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.

Solution:

For $a > 0$, with density $p(z) = (1/2a)\exp(-|z|/a)$, we have that the maximum likelihood estimate is $\widehat{w} = \arg\min_w \|Xw - Y\|_1$.

(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.

Solution:

For uniformly distributed v_i on $[-a, a]$, the density function is $p(z) = \frac{1}{2a}$. A maximum likelihood estimate is any w satisfying $\|Xw - Y\|_\infty \leq a$.

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

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 $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$

Solution:

Note $Xw-Y$ is a vector in $\mathbb{R}^n$, and the i th row has the value $(x_i^\top w - y_i)$. Without loss of generality, let P be vector of any length. By linear algebra, $\|P\|_2$ means $\sqrt{\sum_i P_i^2}$. Also note the identity $P^T P = P \cdot P = \sum_i P_i \cdot P_i = \sum_i P_i^2$. Therefore, $\|P\|_2 = \sqrt{\sum_i P_i^2} = \sqrt{P^T P}$, and thus $\|P\|_2^2 = P^T P = \sum_i P_i^2$. Now substitute $P = Xw - Y$, and we naturally get $\|Xw - Y\|_2^2 = \sum_{i=1}^m (x_i^\top w - y_i)^2$.

2.2. Subspaces of X

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

Solution:

- Rowspace is the **span** (i.e., *the set of all linear combinations*) of the rows of X. Therefore, in this example, it is the subspace of vectors of the form $(1 \cdot x + 4 \cdot y, 2 \cdot x + 5 \cdot y, 3 \cdot x + 6 \cdot y)$ for all x and y.
- Columnspace (a.k.a. Range(X)) is the span of the columns of X. In this example, it is the subspace of vectors of the form $(1 \cdot x + 2 \cdot y + 3 \cdot z, 4 \cdot x + 5 \cdot y + 6 \cdot z)$ for all x, y, and z.
- Nullspace (a.k.a. Null(X)) is the set of vectors v such that $Xv = 0$. In this example, the nullspace is the subspace spanned by $(1, -2, 1)$.
- The matrix X can be reduced to the form $\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$. This matrix has submatrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, which has rank 2. Observe that the third column, $\begin{pmatrix} -1 \\ 2 \end{pmatrix}$, is in the columnspace of this first submatrix.

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.

Solution:

The matrix $X^\top$ is $\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$. The rows of X are the columns 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$(X)^\perp$.

Solution:

A vector $v \in$ Null(X) if and only if $Xv = 0$, which is true if and only if for every row X_i of X, $\langle X_i, v \rangle = 0$. This is precisely the condition that v is perpendicular to each row of X, which is the stated claim.

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

Solution:

This is seen by applying the previous result to $X^\top$.

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?

Solution:

Yes. Symmetry can be checked by computing the transpose. For any vector u, we have $u^\top X^\top X u = \|Xu\|_2^2 \geq 0$.

(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$(X^\top X)$.

Solution:

We have, Null$(X) =$ Null$(X^\top X)$. Let $v \in$ Null(X). Then, one can check that $X^\top X v = 0$, leading to $v \in$ Null$(X^\top X)$, which proves Null$(X) \subseteq$ Null$(X^\top X)$. For the other direction, let $0 \neq v \in$ Null$(X^\top X)$. Then, $0 = v^\top X^\top X v = \|Xv\|_2^2$, which implies $v \in$ Null(X). Therefore, Null$(X^\top X) \subseteq$ Null(X), which finishes the proof.

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

Solution:

No, this isn't always the case. Since Null$(X) =$ Null$(X^\top X)$ (see the answer to the previous question), the matrix $X^\top X$ is not invertible if X has a non-empty nullspace.

(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.

Solution:

If X is "short and wide", it has a non-empty nullspace. Therefore, $X^\top X$ is not invertible.

(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.)

Solution:

$X^\top X$ is symmetric, and previous parts, we have rowspace$(X^\top X) =$ columnspace$((X^\top X)^\top) =$ columnspace$(X^\top X)$. By previous parts again, we have: rowspace$(X^\top X) =$ Null$(X^\top X)^\perp =$ Null$(X)^\perp =$ rowspace(X).

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 from problem 2a of HW 0 that $\mathbb{E}[(Y - \eta(X))^2] = \mathbb{E}[\mathbb{E}[(Y - \eta(X))^2|X = x]]$ (The "Tower Rule").

Solution:

To determine the best $\eta(X)$, we compute the derivative of hint with respect to $\eta(X)$ and set it to zero, as below.

$$\begin{aligned}0 &= \frac{d}{d\eta(X)}\mathbb{E}[(Y - \eta(X))^2|X = x] \\ &= \mathbb{E}[\frac{d}{d\eta(X)}(Y - \eta(X))^2|X = x] \\ &= \mathbb{E}[-2(Y - \eta(X))|X = x] \\ &= -2\mathbb{E}[Y|X = x] + 2\eta(X)\end{aligned}$$

Rearranging, we conclude that the optimal function $\eta(x)$ is $\mathbb{E}[Y|X = x]$.

(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.

Solution:

Let $\eta(x) = \theta$, $\hat{f}_D(x) = \hat{\theta}$ and $\mathbb{E}[\hat{f}_D(x)] = \theta^*$. Note that (given some distribution D) θ and θ^* are numbers and hence $\mathbb{E}[\theta] = \theta$ and $\mathbb{E}[\theta^*] = \theta^*$.

$$\begin{aligned}\mathbb{E}[(\eta(x) - \hat{f}_D(x))^2] &= \mathbb{E}[(\theta - \hat{\theta})^2] \\ &= \mathbb{E}[((\theta - \theta^*) + (\theta^* - \hat{\theta}))^2] \\ &= (\theta - \theta^*)^2 + 2(\theta - \theta^*)\mathbb{E}[\theta^* - \hat{\theta}] + \mathbb{E}[(\theta^* - \hat{\theta})^2] \\ &= (\theta - \theta^*)^2 + \mathbb{E}[(\theta^* - \hat{\theta})^2]\end{aligned}$$

Note that we can do the last step because $\mathbb{E}[\hat{\theta}] = \theta^*$.
The right term is the variance and the left term is the bias squared.

As complexity of F goes up, the bias is decreasing, while the variance is increasing. Thus, we want to find the sweet spot that both of them are reasonably low. This is called bias-variance tradeoff.

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.

Solution:

We first give the proof using “matrix” notation. The objective function can be expressed as

$$\begin{aligned} f(\omega) &= \|X\omega - y\|_2^2 + \lambda\omega^\top D\omega \\ &= (X\omega - y)^\top(X\omega - y) + \lambda\omega^\top D\omega \\ &= (X\omega)^\top X\omega - (X\omega)^\top y - y^\top X\omega + y^\top y + \lambda\omega^\top D\omega \\ &= \omega^\top X^\top X\omega - 2\omega^\top X^\top y + y^\top y + \lambda\omega^\top D\omega \\ &= \omega^\top(X^\top X + \lambda D)\omega - 2\omega^\top X^\top y + y^\top y \end{aligned}$$

The gradient of f is

$$\begin{aligned} \nabla f(\omega) &= \nabla_\omega(\omega^\top(X^\top X + \lambda D)\omega - 2\omega^\top X^\top y + y^\top y) \\ &= \nabla_\omega(\omega^\top(X^\top X + \lambda D)\omega) - 2\nabla_\omega(\omega^\top X^\top y) + \nabla_\omega(y^\top y) \\ &= 2(X^\top X + \lambda D)\omega - 2X^\top y \end{aligned}$$

Here note that $X^\top X + \lambda D$ is a symmetric matrix, which explains the factor 2 in the gradient term. Setting the gradient $\nabla f(\omega)$ to zero, we can conclude that

$$(X^\top X + \lambda D)\widehat{\omega}_{\text{general}} = X^\top y$$

If $X^\top X + \lambda D$ is full rank then we can get a unique solution:

$$\widehat{\omega}_{\text{general}} = (X^\top X + \lambda D)^{-1}X^\top y$$

Since D is already given to be a diagonal matrix with strictly positive entries on the diagonal, any strictly positive λ will make the matrix $X^\top X + \lambda D$ invertible.

Solution:

We now give a solution in the "coordinate" form. The objective, when written in coordinate form, is $f(\omega) = \sum_{i=1}^{n}(y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2$. As in the previous proof, we first simplify it as follows and then set it zero:

$$\begin{aligned}
\nabla_\omega \left[\sum_{i=1}^{n}(y_i - x_i^\top \omega)^2 + \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2\right] &= \nabla_\omega \sum_{i=1}^{n}(y_i - x_i^\top \omega)^2 + \nabla_\omega \lambda \sum_{i=1}^{d} D_{ii}\omega_i^2 \\
&= \sum_{i=1}^{n} \nabla_\omega (y_i - x_i^\top \omega)^2 + 2\lambda D\omega \\
&= -\sum_{i=1}^{n} 2x_i(y_i - x_i^\top \omega) + 2\lambda D\omega \\
&= -\sum_{i=1}^{n} 2x_i y_i + \sum_{i=1}^{n} 2x_i x_i^\top \omega + 2\lambda D\omega \\
&= -2\sum_{i=1}^{n} x_i y_i + 2\left(\sum_{i=1}^{n} x_i x_i^\top + \lambda D\right)\omega \\
&= 0 \quad \text{(set it to be 0)}
\end{aligned}$$

$$\widehat{\omega}_{\text{general}} = \left(\sum_{i=1}^{n} x_i x_i^\top + \lambda D\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right)$$

Note that, as expected, this exactly matches the answer we got from the previous approach (because x_i's are all the rows of X, and therefore $\sum_i x_i y_i = X^\top y$, and $\sum_i x_i x_i^\top = X^\top X$).

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?

Solution:

As can be seen from the formula $\widehat{\omega} = (X^\top X)^{-1} X^\top y$, doubling y doubles ω as well. This makes sense intuitively as well because if the observations are scaled up, the model should also be.

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

Solution:

As can be seen from the formula $\widehat{\omega} = (X^\top X)^{-1} X^\top y$, doubling X halves ω. This also makes sense intuitively because the error we are trying to minimize is $\|X\omega - y\|_2^2$, and if the X has doubled, while y has remained unchanged, then ω must compensate for it by reducing by a factor of 2.

(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.

Solution:

The solution is $\widehat{\omega} = (X^\top X + \lambda I)^{-1} X^\top y$. We already saw in a previous part that $X^\top X$ is always positive semidefinite, that is, its eigenvalues are at least zero. Adding λI, where $\lambda > 0$, ensures that $X^\top X + \lambda I$ is in fact positive *definite*. This helps us have a good condition number.