ECE/CS/ME 539 – Fall 2024 — Activity Solution 13

1 Minimizing Sum of Squared Differences

1.1 Analytic Solution for Optimal b

To find the analytic solution for the optimal value of b that minimizes $\sum_i (x_i - b)^2$, we set the derivative with respect to b equal to zero:

$$\begin{aligned}\frac{d}{db}\left[\sum_i (x_i - b)^2\right] &= -2\sum_i (x_i - b) = 0\\ \sum_i x_i - nb &= 0\\ \sum_i x_i &= nb\\ b &= \frac{1}{n}\sum_i x_i\end{aligned}$$

Therefore, the optimal value of b is the arithmetic mean of the data points.

1.2 Relation to Normal Distribution

This problem and its solution relate to the normal distribution in several ways:

1. **Likelihood Function:** The sum of squared differences is directly related to the likelihood function of the normal distribution. For a normal distribution with mean μ and variance σ^2, the likelihood function is:

$$L(\mu, \sigma^2 | x_1, ..., x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Taking the logarithm and ignoring constant terms, we get:

$$\log L(\mu, \sigma^2 | x_1, ..., x_n) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Maximizing this log-likelihood is equivalent to minimizing $\sum_{i=1}^{n}(x_i - \mu)^2$, which is our original problem with b playing the role of μ.

2. **Maximum Likelihood Estimator:** The arithmetic mean (our solution for b) is the maximum likelihood estimator for the mean of a normal distribution. This can be shown by setting the derivative of the log-likelihood with respect to μ to zero:

$$\frac{\partial}{\partial \mu} \log L = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

Solving this equation gives us $\mu_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$, which is identical to our solution for b.

1.3 Changing the Loss Function

If we change the loss to $\sum_i |x_i - b|$, we're minimizing the sum of absolute deviations. In this case, the optimal solution for b is the median of the data points. Here's a proof:

Let m be the median of the data points $\{x_1, \ldots, x_n\}$.

For any b, consider the difference:

$$D(b) = \sum_i |x_i - b| - \sum_i |x_i - m|$$

We can rewrite this as:

$$D(b) = \sum_{i:x_i<b} (b - x_i) + \sum_{i:x_i>b} (x_i - b) - \sum_{i:x_i<m} (m - x_i) - \sum_{i:x_i>m} (x_i - m)$$

The key is to consider the derivative of $D(b)$ with respect to b:

$$\frac{d}{db} D(b) = |\{i : x_i < b\}| - |\{i : x_i > b\}|$$

At $b = m$, by the definition of median, this derivative is zero (or switches from negative to positive if n is even). This shows that m minimizes $D(b)$, and thus minimizes $\sum_i |x_i - b|$.

2 Equivalence of Affine and Linear Functions

To prove that affine functions expressed as $\mathbf{x}^\top \mathbf{w} + b$ are equivalent to linear functions on $(\mathbf{x}, 1)$:

An affine function $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} + b$ can be rewritten as:

$$f(\mathbf{x}) = [\mathbf{x}^\top, 1] \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$$

This is equivalent to a linear function $g([\mathbf{x}; 1]) = [\mathbf{x}; 1]^\top [\mathbf{w}; b]$ on the augmented vector $(\mathbf{x}, 1)$.

Thus, every affine function on $\mathbf{x}$ can be expressed as a linear function on $(\mathbf{x}, 1)$, and vice versa, proving their equivalence.

3 Full Rank Condition in Linear Regression

In the context of linear regression, we consider the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is the vector of responses, $\mathbf{X}$ is the design matrix, $\boldsymbol{\beta}$ is the vector of coefficients, and $\boldsymbol{\epsilon}$ is the error term.

3.1 Consequences of Rank Deficiency

If $\mathbf{X}^\top\mathbf{X}$ does not have full rank:

1. **Non-unique solutions:** The normal equations $(\mathbf{X}^\top\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y}$ will have infinitely many solutions. This is because the null space of $\mathbf{X}^\top\mathbf{X}$ is non-trivial, meaning there exist non-zero vectors $\mathbf{v}$ such that $(\mathbf{X}^\top\mathbf{X})\mathbf{v} = \mathbf{0}$. If $\boldsymbol{\beta}_0$ is a solution, then $\boldsymbol{\beta}_0 + \alpha\mathbf{v}$ is also a solution for any scalar α.

2. **Undefined least squares estimator:** The least squares estimator $\boldsymbol{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is not uniquely defined because $(\mathbf{X}^\top\mathbf{X})^{-1}$ does not exist when $\mathbf{X}^\top\mathbf{X}$ is not full rank.

3.2 Fixing Rank Deficiency

Adding small Gaussian noise to $\mathbf{X}$ can help resolve the rank deficiency: Let $\tilde{\mathbf{X}} = \mathbf{X} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a matrix of independent Gaussian noise with mean 0 and small variance σ^2.

See 3.3 if you are interested in math details.

3.3 Expected Value After Adding Noise

Let $\boldsymbol{\varepsilon}$ be the noise matrix with i.i.d. Gaussian entries of mean 0 and variance σ^2. We can derive the expected value of $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$:

$$
\begin{aligned}
E[\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}] &= E[(\mathbf{X}+\boldsymbol{\varepsilon})^\top(\mathbf{X}+\boldsymbol{\varepsilon})] \\
&= E[\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\top\mathbf{X} + \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] \\
&= \mathbf{X}^\top\mathbf{X} + E[\mathbf{X}^\top\boldsymbol{\varepsilon}] + E[\boldsymbol{\varepsilon}^\top\mathbf{X}] + E[\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] \\
&= \mathbf{X}^\top\mathbf{X} + \mathbf{0} + \mathbf{0} + n\sigma^2\mathbf{I} \\
&= \mathbf{X}^\top\mathbf{X} + n\sigma^2\mathbf{I}
\end{aligned}
$$

Where n is the number of rows in $\mathbf{X}$, and $\mathbf{I}$ is the identity matrix.

The added term $n\sigma^2\mathbf{I}$ ensures that the expected $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ has full rank, as it adds a positive value to each diagonal element.

Fixing Rank Deficiency Details

To address the issue of rank deficiency in the design matrix $\mathbf{X}$, we can add a small amount of coordinate-wise independent Gaussian noise to all entries of $\mathbf{X}$. Let's explore this approach mathematically:

Let $\tilde{\mathbf{X}} = \mathbf{X}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a matrix of independent Gaussian noise with each entry $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, and σ^2 is small.

3.3.1 Effect on Rank

Theorem 1 *With probability 1, the matrix $\tilde{\mathbf{X}}$ has full rank.*

Proof: Let $\mathbf{X}$ be an $n \times p$ matrix. The statement that $\tilde{\mathbf{X}}$ is rank deficient is equivalent to the existence of a non-zero vector $\mathbf{v} \in \mathbb{R}^p$ such that $\tilde{\mathbf{X}}\mathbf{v} = \mathbf{0}$.

This means:

$$(\mathbf{X} + \boldsymbol{\varepsilon})\mathbf{v} = \mathbf{0}$$

$$\boldsymbol{\varepsilon}\mathbf{v} = -\mathbf{X}\mathbf{v}$$

For any fixed $\mathbf{v}$, the left-hand side $\boldsymbol{\varepsilon}\mathbf{v}$ is a vector of independent Gaussian random variables (since linear combinations of independent Gaussian variables are Gaussian). The probability that this exactly equals $-\mathbf{X}\mathbf{v}$ is zero.

There are uncountably many possible $\mathbf{v}$, but the union of countably many zero-probability events still has probability zero. Therefore, the probability that there exists any $\mathbf{v}$ satisfying this equation is zero.

Thus, with probability 1, no such $\mathbf{v}$ exists, meaning $\tilde{\mathbf{X}}$ has full rank. □

3.3.2 Effect on Condition Number

While adding noise ensures full rank with probability 1, it's also important to consider how it affects the condition number of the matrix, which measures how close a matrix is to being singular.

Lemma 2 *Adding small Gaussian noise tends to improve (i.e., reduce) the condition number of ill-conditioned matrices.*

Proof:[Sketch of proof] The condition number $\kappa(\mathbf{A})$ of a matrix $\mathbf{A}$ is defined as:

$$\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values of $\mathbf{A}$.

For an ill-conditioned matrix, $\sigma_{\min}$ is very close to zero. Adding Gaussian noise increases all singular values by an expected amount related to the noise variance. This increase has a much larger relative effect on small singular values than on large ones, thus reducing the ratio $\sigma_{\max}/\sigma_{\min}$ and improving the condition number. □