COMP 562 – Lecture 4

Linear Regression -- Matrix Form

A general multiple-regression model can be written as

$$y|\mathbf{x} = \beta_0 + \sum_j x_j\beta_j + \epsilon, \quad \epsilon \sim \mathcal{N}\left(, 0, \sigma^2\right)$$

Which is equivalent to

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \quad for\, i = 1, \ldots, N$$

In matrix form, we can rewrite this model as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Np} \end{bmatrix}_{N\times p+1} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix}_{p+1\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}_{N\times 1}$$

This can be rewritten more simply as:

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

Linear Regression -- Closed Form Solution for β

Remember that maximizing log-likelihood is equivalent to minimizing **RSS** or **MSE**

$$RSS = \sum_{i=1}^{N}\left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j)\right)^2 = \sum_{i=1}^{N}(e_i)^2 = \mathbf{e_i}^T\mathbf{e_i} = \begin{bmatrix} e_1 & e_2 & \ldots & e_N \end{bmatrix}_{1\times N}\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}_{N\times 1}$$

$$\begin{aligned} RSS &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \\ &= (\mathbf{y}^T - \beta^T\mathbf{X}^T)(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta \\ &= \mathbf{y}\mathbf{y}^T - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \end{aligned}$$

Where this development uses the fact that the transpose of a scalar is the scalar i.e. $\beta^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X}\beta$

To find the β that minimizes RSS, we solve the following equation:

$$\nabla_\beta \, RSS = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$

The corresponding solution to this linear system of equations is called the **ordinary least squares** or **OLS** solution

$$\hat{\beta} = \beta^{\text{OLS}} = \beta^{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Linear Regression -- Closed Form Solution for σ^2

Recall log-likelihoood function

$$\log \mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N}\left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j)\right)^2\right]$$

Which can be written in matrix form

$$\log \mathcal{L}(\beta|\mathbf{y}, \mathbf{x}) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

Taking derivative and equating it to zero yields

$$(\sigma^2)^{\text{MLE}} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\beta^{\text{MLE}})^T(\mathbf{y} - \mathbf{X}\beta^{\text{MLE}}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - (\beta_0^{\text{MLE}} + \sum_j x_{i,j}\beta_j^{\text{MLE}})\right)^2$$

Please verify $(\sigma^2)^{\text{MLE}}$ at home

- **Overfitting**: Model every minor variation in the input using highly flexible complex models
  - High variance and low bias
- **Underfitting**: Simple model that is unable to capture the true relationships in given data
  - Low variance and high bias
- **Model Selection**: Picking the right model from a variety of models of different complexity

Q: Which model overfits/underfits the data?

Bias-Variance Tradeoff

The **mean quared errors** or **MSE** may be decomposed into **bias** and **variance** components:

$$\underbrace{\mathbb{E}(y-\hat{y})^2}_{\textbf{MSE}} = \underbrace{(\mathbb{E}(\hat{y})-y)^2}_{\textbf{Bias}^2} + \underbrace{\mathbb{E}\left[(\hat{y}-\mathbb{E}(\hat{y}))^2\right]}_{\textbf{Variance}} + \underbrace{\sigma_e^2}_{\textbf{Irreducible Error}}$$

Ill-Posed Problems

Q: What happens if you are solving a linear system $Ax = y$ and there are more unknowns than equations?

In our setting -- N samples, P features -- linear regresion is ill-posed if $P > N$

Another example of ill-posed linear regression problem arises when we have two copies of the same predictors

This is a problem even if $P < N$

Ridge Regression

Adding that penalty to linear regression log-likelihood yields **ridge regresion**

$$\log \mathcal{L}(\beta_0, \beta | \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j) \right)^2 \right] - \underbrace{\frac{\lambda}{2}\sum_j \beta_j^2}_{\text{ridge penalty}}$$

All those sums can get cumbersome, so we will use norms

1. ℓ_2 norm $\|\mathbf{x}\| = \sqrt{\sum_i x_i^2}$
2. ℓ_1 norm $\|\mathbf{x}\|_1 = \sum_i |x_i|$

$$\log \mathcal{L}(\beta | \mathbf{y}, \mathbf{x}) = -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2 \underbrace{-\frac{\lambda}{2}\|\beta\|^2}_{\text{ridge penalty}} + \text{const.}$$

Ridge Regression -- Computing Gradients

$$\log \mathcal{L}(\beta|\mathbf{y}, \mathbf{x}) = -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\beta\|^2 \underbrace{-\frac{\lambda}{2}\|\beta\|^2}_{\text{ridge penalty}} + \text{const.}$$

Computing the gradient and setting it to zero

$$\nabla_\beta \log \mathcal{L}(\beta|\mathbf{y}, \mathbf{x}) = \frac{1}{\sigma^2} X^T(\mathbf{y} - \mathbf{X}\beta) - \lambda\beta = 0$$

yields

$$\beta^{\text{MLE}} = (\mathbf{X}^T\mathbf{X} + \lambda\sigma^2\mathbf{I}_N)^{-1}(\mathbf{X}^T\mathbf{y})$$

Where $\mathbf{I}_N$ is the identity matrix of size N

Contrast this to closed form solution of linear regression

$$\beta^{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Ridge Regression -- Computing Gradients

The bias/intercept coefficient β_0 is typically not regularized in a linear regression

A regularized β_0 (shrinked) may us from prevent finding the correct relationship

$$\nabla \log \mathcal{L}(\beta_0, \beta, \sigma^2|\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \sum_{i=1}^N -\frac{1}{\sigma^2}\left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j)\right)(-1) \\ \sum_{i=1}^N -\frac{1}{\sigma^2}\left(y_i - (\beta_1 + \sum_j x_{i,j}\beta_j)\right)(-x_{i,1}) - \lambda\beta_1 \\ \vdots \\ \sum_{i=1}^N -\frac{1}{\sigma^2}\left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j)\right)(-x_{i,p}) - \lambda\beta_p \end{bmatrix}$$

Note that β_0 is **not** regularized

Remember our closed form solution for ridge regression

$$\beta^{\text{MLE}} = (\mathbf{X}^T\mathbf{X} + \lambda\sigma^2\mathbf{I}_N)^{-1}(\mathbf{X}^T\mathbf{y})$$

Updating our closed form solution without regularizing β_0 will yeild

$$\beta^{\text{MLE}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\sigma^2\begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}_{N\times N}\right)^{-1}(\mathbf{X}^T\mathbf{y})$$