1

STUDY GUIDE

1.1 Linear Regression Models

Suppose we have an input vector $\mathbf{X}^\top = (X_1, X_2, \ldots, X_p)$ and we want to predict a real-valued output Y. The linear regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \tag{1.1}$$

Typically we have a set of training data $(x_1, y_1) \ldots (x_N, y_N)$ from which we can estimate the parameters β by minimizing the error function which we will define next. Each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ represents a vector of feature measurements for the ith case. The most popular estimation method is *least squares*, in which we pick the coefficients β to minimize the residual sum of squares

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} \left(y_i - f\left(x_i\right)\right)^2 \\ &= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 . \end{aligned} \tag{1.2}$$

To minimize (1.2), we rewrite β by appending $\beta_0 = 1$ at its first element for convenience. Then, we rewrite the residual sum-of-squares as

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta). \tag{1.3}$$

Expanding the above equality, we could obtain

$$\begin{aligned} \mathrm{RSS}(\beta) &= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\beta - \beta^\top\mathbf{X}^\top\mathbf{y} + \beta^\top\mathbf{X}^\top\mathbf{X}\beta \\ &= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\beta + \beta^\top\mathbf{X}^\top\mathbf{X}\beta \end{aligned} \tag{1.4}$$

Differentiating with respect to β we obtain

$$\begin{aligned} \frac{\partial \mathrm{RSS}}{\partial \beta} &= -2\mathbf{y}^\top\mathbf{X} + 2\mathbf{X}^\top\mathbf{X}\beta \\ \frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta^\top} &= 2\mathbf{X}^\top\mathbf{X}. \end{aligned} \tag{1.5}$$

Since the Hessian $\mathbf{X}^\top\mathbf{X}$ is positive definite, we set the first derivative to zero

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{1.6}$$

to obtain the unique solution

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{1.7}$$

The predicted values at an input vector x_0 are given by $\hat{f}(x_0) = (1 : x_0)^\top\hat{\beta}$; the fitted values at the training inputs are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \tag{1.8}$$

where $\hat{y}_i = \hat{f}(x_i)$.

1.2 Conditioning Gaussians

1.3 Ridge Regression / L1 regularized Linear Regression

1.4 L1/L2 Regularization

1.5 Multivariate Gaussian

1.6 Gaussian Discriminant Analysis

1.7 Exponential Family

1.8 Solving an Unconstrained Quadratic

1.9 SVD, Cholesky, Eigen/Spectral Decompositions

1.10 Generalized Linear Models