Statistical Machine Learning

Linear regression

Linear regression

- Linear regression, also called the method of least squares, is an old topic, dating back to Gauss in 1795 (he was 18!).
- Linear regression is a simple approach to supervised learning. It assume that the dependence of Y on $X_1, \ldots, X_p$ is linear.
- True regression functions are never linear!
- although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.
- we all know simple linear regression model, let's directly talk about multiple linear regression

Multiple linear regression

- Suppose we are considering $Y \in \mathbb{R}^n$ as a function of multiple predictors $X_1, \ldots, X_p \in \mathbb{R}^n$. We collect these predictors into columns of predictor matrix (design matrix) $X \in \mathbb{R}^{n \times p}$. Assume that $X_1, \ldots, X_p$ are linearly independent, so that $\text{rank}(X) = p$.
- We write

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon,$$

  where β_j's are the coefficients. ε's are independent errors with mean 0 and and sd σ
- we estimate $\beta = (\beta_0, \ldots, \beta_p)$ by least square:

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \sum_i (Y_i - \beta_0 - \beta_1 X_{1i} - \ldots - \beta_p X_{pi})^2$$

- This gives the minimizer $\hat{\beta}$ and the fitted value

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_p X_{pi},$$

Assessing the Accuracy

- of the Coefficient Estimates
  - Confidence intervals
- of the Model
  - R^2 or fraction of variance explained

$$R^2 = 1 - SSE/SST$$

where $SSE = \sum_i (y_i - \hat{y}_i)^2$ and $SST = \sum_i (y_i - \bar{y}_i)^2$

Deciding on the important variables

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- stepwise selection such as forward and backward selection
- Never use the ordinary R^2 to compare models
- Adjusted R^2, Generalized Cross validation (GCV), Akaike information criterion (AIC), Bayesian information criterion (BIC) and many others

Assessing the Accuracy of the estimated regression model

- Split the data into train and test sets
- Train models using the training set
- Compute MSEs on the test set
- Choose the model with the smallest MSE (over the test set)

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated
  - a balanced design:
  - each coefficient can be estimated and tested separately
  - interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible.
- correlations amongst predictors cause problems
  - the variance of all coefficients tends to increase, sometimes dramatically
  - interpretations become hazardous: when X_j changes, everything else changes.
- claims of causality should be avoided for observational data.
  - drownings vs icecream sales
  - married men live longer than single men
  - chocolate causes xxxxxx

Potential problems

- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.

Shortcomings of regression

- Predictive ability: the linear regression fit often does not predict well, especially when p (the number of predictors) is large

(Important to note that is not even necessarily due to nonlinearity in the data! Can still predict poorly even when a linear model could fit well)

- Interpretative ability: linear regression "freely" assigns a coefficient to each predictor variable. When p is large, we may sometimes seek, for the sake of interpretation, a smaller set of important variables

Hence we want to "encourage" our fitting procedure to make only a subset of the coefficients large, and others small or even better, zero

Generalizations of the Linear Model

In much of the rest of this course, we discuss methods that expand the scope of linear models and how they are fit:

- Classification problems: logistic regression, support vector machines
- Non-linearity: kernel smoothing, splines and generalized additive models; nearest neighbor methods.
- Interactions: Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- Regularized fitting: Ridge regression and lasso