Supervised Learning

Key points

- Regression and classification.
- Least Squares (LS) vs K-Nearest Neighbors (KNN)
- Curse of Dimensionality
- Central Limit Theorem
- Assumptions of Linear Regression

Main types of models

- Regression: predict quantitative outputs
- Classification: predict qualitative outputs

Least Squares

LS vs KNN: two extreme examples of models in the spectrum.

Linear model: assumed structure

KNN: non-parametric. Assumes no structure.

> 💡 The linear model makes huge assumptions about structure and yields stable but possibly inaccurate predictions.

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j.$$

- Linear model in vector form

$$\hat{Y} = X^T \hat{\beta},$$

  - X^T denotes vector or matrix transpose. X being a column vector ($m * 1$-matrix with 1 column)
  - Y hat is a scalar
- General linear model form:

Here we are modeling a single output, so $\hat{Y}$ is a scalar; in general $\hat{Y}$ can be a K–vector, in which case β would be a $p \times K$ matrix of coefficients. In the $(p+1)$-dimensional input–output space, $(X, \hat{Y})$ represents a hyperplane. If the constant is included in X, then the hyperplane includes the origin and is a subspace; if not, it is an affine set cutting the Y-axis at the point $(0, \hat{\beta}_0)$. From now on we assume that the intercept is included in $\hat{\beta}$.

Least squares

- Core: pick the coefficient beta to minimize residual sum of squares RSS
  - stable - result in low variance but high bias (**heavy assumption** on the linear decision boundary)

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

KNN

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

- non-parametric model
- n >> p
- Uses metric Euclidean distance - we find the k observations with x_i closest to x in input space and average their responses, assign y hat to y if the distance of x is close to x
- Unstable - high variance but low bias (no assumption on shape of data / can single out various shape of subregions)
- Tend to overfit

Curse of dimensionality

- Feature selection / dimensionality
  - dimensions = features or attributes
  - with more dimensions, you need large amount of data whose features are abundant in each feature/dimension (otherwise scarcity of feature/data points will confuse the model); otherwise, large set of features are not useful. Only save the informative features.
  - this is why we need to do feature selection / dimensionality reduction

> 💡 if we wish to be able to estimate such functions with the **same accuracy as** function in low dimensions, then we need the size of our training set to grow exponentially as well.

Bias-variance decomposition - MSE

$$
\begin{aligned}
\mathrm{MSE}(x_0) &= \mathrm{E}_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\
&= \mathrm{E}_{\mathcal{T}}[\hat{y}_0 - \mathrm{E}_{\mathcal{T}}(\hat{y}_0)]^2 + [\mathrm{E}_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\
&= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0). \qquad (2.25)
\end{aligned}
$$

two components: variance + squared bias

Statistical Models, Supervised Learning and Function Approximation

Central Limit Theorem

If you repeatedly sample a random variable a large number of times, the distribution of sample mean will approach a normal distribution regardless of the initial distribution.

Assumptions of Linear Regression

1. **X and Y has linear relationship. <MUST MEET>**
   a. **diagnose**: residual on y-axis , predicted value of y on x-axis; points should be symmetrically distributed around the line
2. **Residuals are independent. <MUST MEET>**
   a. often violated if time series, the successive residual tend to be positive.
   b. **diagnose**:
      i. check if data is collected in a sequence. whether observe a pattern
      ii. using y vs x; residuals vs predicted values
3. **The residuals are normally distributed. <NICE TO HAVE> Multivariate Normal - lienar combination of variables should follow a normal distribution**
   a. CLT gives us the parameter estimates and predicted values of the dependent variable are approximately normally distributed even if the residuals are not.

b. Violation of normality

c. **diagnose:**

   i. **histogram** (count vs variable) AND check **quantile-quantile(Q-Q) plot** of the residuals. if the distribution of residuals is normal then the shape should follow a linear line.

4. **Residuals should have equal variance (homoscedasticity) <less important>**

   a. the residuals should have approximately equal variances.

   b. **diagnose:** check residual plot