Chapter 2 Statistical Learning

Statistical Learning

2.1.1 Why estimate F?

1. To make prediction
2. To understand the relationship:
   sometimes, we care less about prediction and more about insight
   -How does the salary depend on the experience?
   -Which feature matter most?
   Estimate the F help explain how X affect Y
3. Because the noise exist
   Real world data is messy even with the same X, output vary. So, instead of learning noise, we aim for a function that capture the pattern, not randomness.

How can we estimate the F?
Step 1:
Choose a form of F (2 Approaches)
-Parametic method
Asssume a specific shape for F
Ex: Linear regression
Y = B0 + B1X
Pros:
-Simple - fast - easy to interpret
Cons:
-can be too regid (simple) (Underfitting)
-Non parametic method
do not asssum a fixed form
Ex: kNN, neural network, Decision tree, Random forest
Pros:
-flexible - can model complex pattern
Cons:
-need more data - can overfitting
Step 2:
Define s lose function:
This measures how wrong our estimate is
Ex: -Mean Square Error (regression) - Miss-entropy loss (Classification)
Step 3: (Learn from data)
We find F^ that minimize the loss
This can be done by gradient Descent and Closing-form Solution

2.2 Assessing Model Accuracy
2.2.1 Assessing the model accuracy
A mode that work only on training data is useless
we care about: how well the model performs on unseen data.
→ Always evaluate models using *test error*, not *training error*
2.2.2 Bias-Variance trade-off
Why test error exist?
Even the best model make mistake because:
-Wrong assumption (Bias)
-Sensitivity tp data (Variances)
-noise (irreducible error)
Bias = Error caused by oversimplifying the real relationship.
characteristic:
-Model too rigid
-Misses imported pattern
-lead to underfitting
Ex: using a straight line to fit a curve pattern.
Variance = error caused by model sensitivity to training data.
Characteristic:
-Model change a lot with small data change
-fit noise
-lead to overfitting
Irreducible error:
-Error caused by noise in data
-cannot be eliminated
-set a lower bound on test error
Mathematic composition
Expected test MSE = Bias** 2 + Variance + Irreducible error
2.2.3 Classification setting
Classification error rate
Error rate = incorrect prediction/Total prediction
Model sensitivity in classification
+Simple classifier
-High bias
-Low variance
+Complex classifier
-low bias
-High variance
Ex:
-K-NN with large K → Underfitting
-k-NN with small K → Overfitting