Lecture 2 Statistical Machine Learning

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

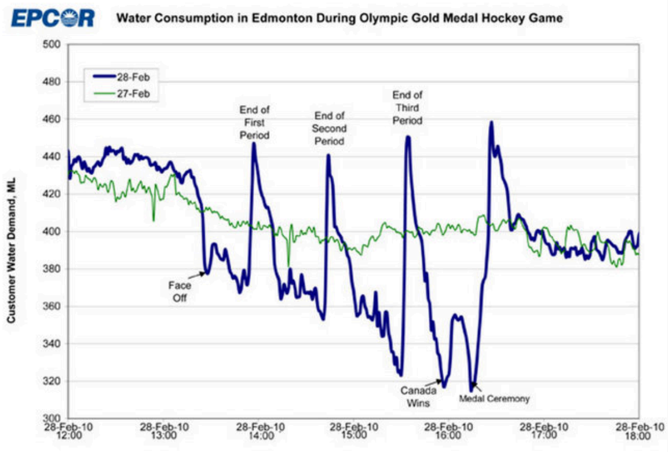
Department of Electrical and Computer Engineering
University of Alberta

January 13, 2021

Outline

Seeing the data

- They say a picture is worth 10,000 words



Water Use (Toilet Flushes) in Vancouver 2010 final Canada vs. USA

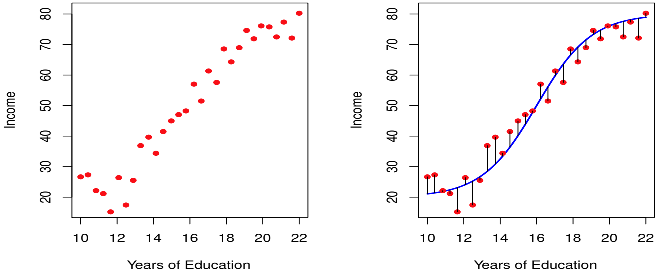
Statistical Machine Learning

- Given response Y_i and covariates $\boldsymbol{X}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})^T$, we model the relationship

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i,$$

where f is an unknown function and ε is random error with mean zero.

- A Simple example



Estimate or learn the relationship

- A typical supervised learning task is to estimate the relationship f, or using data to learn f. Why?
- To make prediction for the response Y for a new value of $\boldsymbol{X}$;
- To make inference on the relationship between Y and $\boldsymbol{X}$, say, which x actually affect Y, positive or negative, linearly or more complicated.
- Prediction Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
- Inference Wish to predict median house price based on 14 variables. Probably want to understand which factors have the biggest effect on the response and how big the effect is.

Estimate or learn the relationship

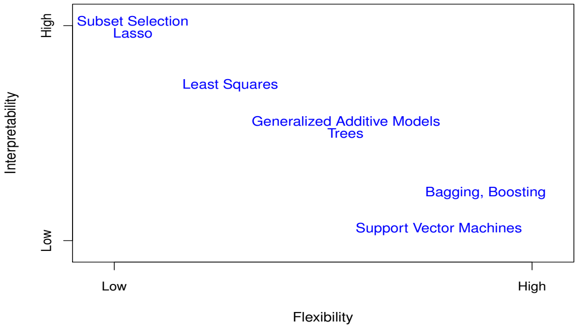
- How to estimate or learn f?
- Parametric methods say, linear regression (Chapter 3)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

  by certain loss function, e.g. ordinary least squares (OLS).
- Nonparametric methods, say, spline expansion (Chapter 5) and kernel smoothing (Chapter 6) methods.
- Nonparametric methods are more flexible but need more data to obtain an accurate estimation.

Tradeoff between accuracy and interpretability

- The simpler, the better — the law of parsimony or Occam's razor.
- A simple method is much easier to interpret, e.g. linear regression model.
- A simple model is possible to achieve more accurate prediction without overfitting. It seems counter intuitive though.



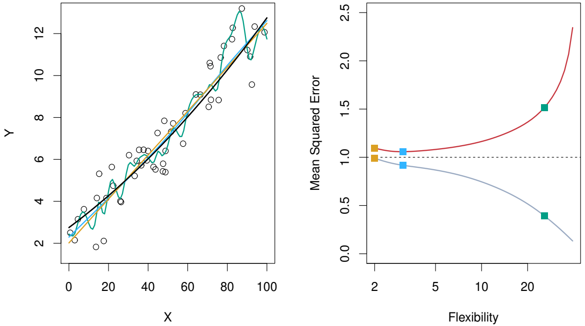
Quality of fit

- A common measure of accuracy is the mean squared error (MSE),

$$\text{MSE} = 1/n \sum_i \left(Y_i - \hat{Y}_i\right)^2,$$

  where $\hat{Y}_i$ is the prediction using the training data.
- In general, we minimize MSE and care how the method works for new data, we call it test data.
- More flexible models could have lower MSE for training data but higher test MSE.

Levels of flexibility



- Black - Truth; Orange - Linear Estimate; Blue - smoothing spline; Green - smoothing spline (more flexible)
- RED - Test MSE; Grey - Training MSE; Dashed - Minimum possible test MSE (irreducible error)

Bias-Variance tradeoff in supervised learning

- There are always two competing forces that govern the choice of learning method i.e. bias and variance.
- Bias refers to the error that is introduced by modeling a real life problem (that is usually extremely complicated) by a much simpler problem.
- The more flexible/complex a method is the less bias it will generally have.
- Variance refers to how much your estimate for f would change by if you had a different training data set.
- Generally, the more flexible a method is the more variance it has.

Bias and Variance tradeoff

- For a new observation Y at $\boldsymbol{X} = \boldsymbol{X}_0$, the expected MSE is

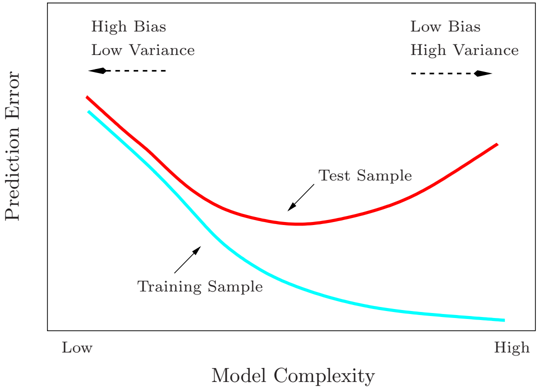
$$E\left[\left(Y - \hat{Y}|\boldsymbol{X}_0\right)^2\right] = E\left[\left(f(\boldsymbol{X}_0) + \varepsilon - \hat{f}(\boldsymbol{X}_0)\right)^2\right] = \text{Bias}^2\left[\hat{f}(\boldsymbol{X}_0)\right] + \text{Var}\left[\hat{f}(\boldsymbol{X}_0)\right] + Var[\varepsilon],$$

  where

$$\text{Bias}\left[\hat{f}(\boldsymbol{X}_0)\right] = E\left[f(\boldsymbol{X}_0) - \hat{f}(\boldsymbol{X}_0)\right],$$

  and the expectation is taken over all training datasets.
- What this means is that as a method gets more complex the bias will decrease and the variance will increase but expected test MSE may go up or down!

Bias and Variance tradeoff



- As a method gets more complex the bias will decrease and the variance will increase but expected test MSE may go up or down!

Summary and Remark

- What is Statistical Machine Learning
- How to assess the accuracy
- Read textbook Chapter 1 and 2
- Do R lab