Introduction to Statistical Learning

(Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor)

quantitative response Y and p different predictors, $X_1, X_2, \ldots X_p$
relationship b/w them expressed as $Y = f(X) + \epsilon$.
ϵ – error term.

Statistical learning – set of approaches for estimating f.
and evaluating the estimates obtained

$\hat{Y} = \hat{f}(X)$

prediction has reducible & irreducible errors
↓
potentially increase the accuracy of $\hat{f}$
using most appropriate statistical learning technique

irreducible error – Y is also a function of ϵ.
∴ variability associated with ϵ also affects the prediction

why irreducible error larger than zero?
it may contain unmeasured variables that are useful for predicting Y
it may also contain unmeasurable variation
(ex – effect of drug on a patient might be dependent on manufacturing variation / patient's well being or mood that day)

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]$$
$$= [f(X) - \hat{f}(X)]^2 + \text{Var}(\epsilon)$$
reducible irreducible

- which predictors are associated with the response?
- Relationship b/w predictors & outcome? (correlation)
- can the relationship be summarised by linear equation, or a complex one?

Estimating f

2 approaches - parametric & non parametric

parametric

steps:

1) assume the functional form or shape of f

ex: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

2) procedure to fit or train the model

ex: estimation of parameters $\beta_0, \beta_1, \ldots \beta_p$

rather than trying to figure out an entire formula, the problem is reduced to find only the parameters $\beta_0, \beta_1, \ldots \beta_p$.

non parametric methods

avoids the assumption of form of f.

but requires huge data to get this estimate of f

ex: using a thin plate spline to fit a data with 3 variables - ~~age~~ seniority, education and income

need to find the right amount of flexibility

flexibility vs interpretability

(good for inference applications)

Supervised vs unsupervised learning

semi supervised - where outcomes are available for only few

Regression vs classification - quantitative vs qualitative

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

expected test MSE:

$$E(y_0 - \hat{f}(x_0)) = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

variance refers to the amount by which $\hat{f}$ would change if training set is different

bias - error introduced by approximating a real life problem

bias variance tradeoff
as model flexibility increases, bias goes down, variance increases

Classification settings

$$\text{error rate} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

I is the indicator saying whether the classification is correct.

$I(y_i \neq \hat{y}_i) = 1$ if $y_i \neq \hat{y}_i$.

$I(y_i \neq \hat{y}_i) = 0$ if $y_i = \hat{y}_i$

Bayes Classifier:
assigning each observation to a particular class which is most likely as per the predictor values.

$$Pr(Y=j \mid X=x_0)$$

ex: in a 2 class problem if $Pr(Y=1 \mid X=x_0) > 0.5 \rightarrow x_0$ belongs to class 1.

Bayes decision boundary - where the prob is exactly 0.5.

Since the classification is done for which the prob is largest, the error rate is $1 - \max_j Pr(Y=j \mid X=x_0)$ at $X=x_0$

overall Bayes error rate is

$$1 - E\left(\max_j Pr(Y=j \mid X)\right)$$

where the expectation averages the prob over all possible values of X.

it is greater than zero because some of the classes overlap.

∴ Bayes error rate is irreducible

k Nearest Neighbours

closest approximation to Bayes classification, to overcome the challenge of unknown conditional probability

k - positive number

identify k points / observations nearest to x_0 - the area represented by N_0

$$P_r(Y=j \mid X=x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j)$$

Exercises

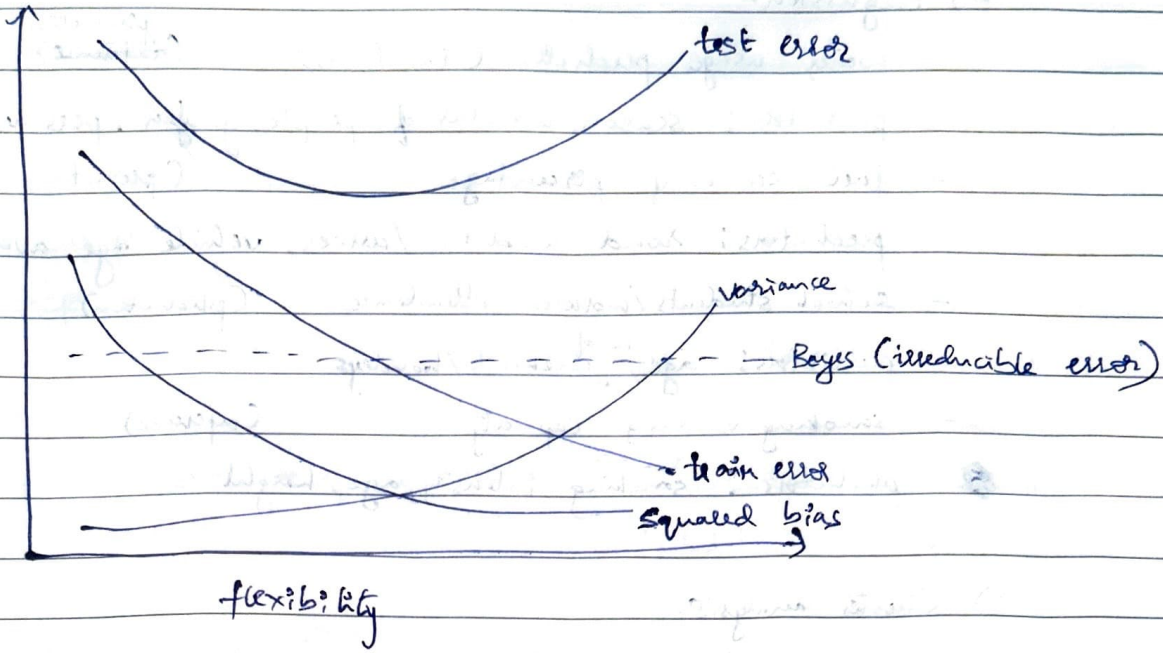
Which performance would be better? Flexible or Inflexible

a) Sample size n is extremely large, predictors p is small
→ flexible method. as there is a chance of overfitting due to large sample size & less predictors

b) p is extremely large, n is small
→ less flexible would be better. else it would overfit
flexible method would fit the noise as well
in small samples, lower variance matters a lot
so that test error (using different data) will be lower

c) Relationship b/w predictors & response is highly non linear
→ flexible model would do better. else it won't be able to capture the non linear pattern of the data

d) variance of the error terms $\sigma^2 = Var(\epsilon)$ is extremely high
→ inflexible would do better. high noise makes it hard to learn the true signal.
high variance should be reduced ⇒ we should use model which has low variance ⇒ inflexible model



Squared bias / MSE reduces as flexibility increases (tighter fit)
variance - as the model is more fit to trained data, it is likely to give lesser accurate result for test data
noise will also be captured in the trained model
Bayes error is constant, irrespective of data and model flexibility
training error reduces as the flexibility increases. But after a point it goes below the irreducible error line, indicating overfitted model
test error - the expected error is given by variance + bias + Bayes error
therefore it is always started at some point above Bayes error curve. decreases as flexibility increases, and then again increases due to variance adding up.

Real life problems that can be solved by

a) Classification
- news consumption (inference)
predictors: age, occupation
- overbought / oversold stock (prediction)
predictors: intrinsic value, CMP, volume
- movie genre (prediction)
predictors: metadata, plot

b) Regression-

- water usage prediction (in Litres) (prediction)
  predictors: season, number of people, garden, pets, vehicles
- fuel economy percentage (prediction)
  predictors: road condition/curves, vehicle age, average speed
- school students/workers attendance (prediction)
  predictors: age, no. of weekend/holidays
- smoking vs lung capacity (inference)
  predictors: smoking intensity, age, height

c) Cluster analysis-

- app users behaviour (power users, casual users, one time users)
  predictors: session duration, feature usage, time of day active
- grouping diseases together in classes based on their features

	Advantages	Disadvantages	Useful when
Flexible Model	can capture more complex & non linear relationships better	more prone to overfit	large training dataset
Inflexible Model			sample size is small, underlying distribution seems linear

If the Bayes decision boundary is highly non linear, smaller K in the KNN algorithm will be more flexible and be able to capture non linearity better.