PQHS 471
Lecture 5:
Fundamentals in Supervised Learning

Starting point

The supervised learning problem:

- Outcome measurement $\mathbf{Y}$ (also called dependent variable, response, target).
- Vector of p predictor measurement $\mathbf{X}$ (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, $\mathbf{Y}$ is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer type/class of tissue sample).
- We have training data (x_1, y_1),...,(x_N, y_N). These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

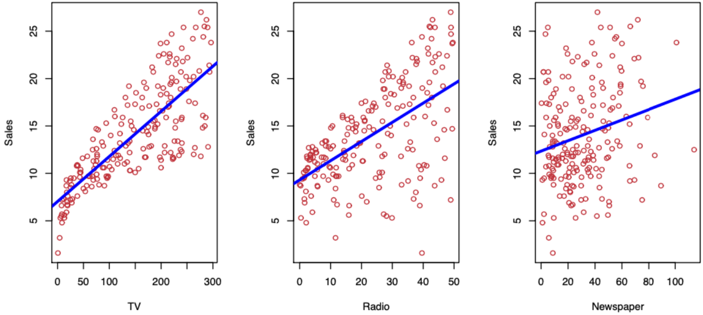
Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in biomedical, banking, IT and finance.
- Supervised learning is a fundamental ingredient in the training of a modern data scientist.

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap — both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on large scale large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".

Supervised learning example



Shown are Sales vs advertising on TV, Radio and Newspaper, with a blue linear-regression line fit separately to each.
Can we predict Sales using these three?
Perhaps we can do better using a model

$$Sales \approx f(TV, Radio, Newspaper)$$

Example notation

Here, Sales is a *response* or *target* that we wish to predict. We generically refer to the response as $\mathbf{Y}$.
TV is a feature, or input, or predictor; we name it X_1.
Likewise name Radio as X_2, and so on.
We can refer to the input vector collectively as

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Now we write our model as

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon$$

where ϵ captures measurement errors and other discrepancies.

Why estimate f ?

- **Prediction**
- **Inference**

Prediction

$\hat{f}$: The estimated f that we want to use.
Once we have an $\hat{f}$, we can make predictions of Y at new points X_{new}

$$\hat{Y} = \hat{f}(X_{new})$$

Here, $\hat{Y}$ represents the resulting prediction of Y.

Prediction

$\hat{f}$: The estimated f that we want to use.
Once we have an $\hat{f}$, we can make predictions of Y at new points X_{new}

$$\hat{Y} = \hat{f}(X_{new})$$

Here, $\hat{Y}$ represents the resulting prediction of Y.
In prediction, $\hat{f}$ can be treated as a *black box*: As long as the prediction is accurate, the exact form of $\hat{f}$ is not the the focus in prediction problems.

Accuracy in Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible Error**
- **Irreducible Error**

Accuracy in Prediction

The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

- **Reducible Error**
- **Irreducible Error**

Reducible: we can potentially improve the accuracy of $\hat{f}$ by using a more proper statistical model/technique.
Irreducible: $\epsilon = Y - f(X)$. The variability associated with ϵ, *cannot* be predicted by X, no matter how well we estimate f.

Reducible vs. Irreducible

Suppose we have a estimate $\hat{f}$ and a set of predictors X, so we have $\hat{Y} = \hat{f}(X)$. Assume both $\hat{f}$ and X are fixed.
It is easy to show that at $X = x$:

$$E(Y - \hat{Y})^2 = E[f(x) + \epsilon - \hat{f}(x)]^2 = [f(x) - \hat{f}(x)]^2 + Var(\epsilon)$$

Reducible vs. Irreducible

Suppose we have a estimate $\hat{f}$ and a set of predictors X, so we have $\hat{Y} = \hat{f}(X)$. Assume both $\hat{f}$ and X are fixed.
It is easy to show that at $X = x$:

$$E(Y - \hat{Y})^2 = E[f(x) + \epsilon - \hat{f}(x)]^2 = [f(x) - \hat{f}(x)]^2 + Var(\epsilon)$$

Now we can see better why the accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

$$\underbrace{[f(x) - \hat{f}(x)]^2}_{\textit{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\textit{Irreducible}}$$

Inference

Beside prediction accuracy, we are often interested in understanding the way that Y is affected as $X_1, X_2, ..., X_p$ change.

- We wish to estimate f, but our goal is not always to predict Y
- Understand the relationship between X and Y
- Understand how Y changes as a function of $X_1, X_2, ..., X_p$
- $\hat{f}$ **can not be regarded as a black box**.

Common inference question #1

Which predictors are associated with the response?

- Often, only a small fraction of the variables are substantially associated with Y.
- Identifying the few *important* predictors can be extermely useful, depending on the application.

Common inference question #2

What is the relationship between the response and each predictor?

- Positive relationship? Negative relationship?
- Association depending on other covariates?

Common inference question #3

For $Y \sim X$: linear equation? More complicated?

- Historically, most methods for estimating f have taken a linear form.
- In some situations, such assumptions are reasonable or even desirable.
- The true relationship can be more complicated.

Common inference question #3

For $Y \sim X$: linear equation? More complicated?

- Historically, most methods for estimating f have taken a linear form.
- In some situations, such assumptions are reasonable or even desirable.
- The true relationship can be more complicated.

Prediction and inference are not mutually exclusive: we will see examples that fall into the prediction setting, the inference setting, or a combination of the two.

Advertising data example questions

One may be interested in answering **inference questions** such as:

- Which media contribute to sales?
- Which media generate the biggest boost in sales?
- How much increase in sales is associated with a given increase in TV advertising?

How to estimate f

- **Parametric methods**
- **Non-parametric methods**

Parametric methods

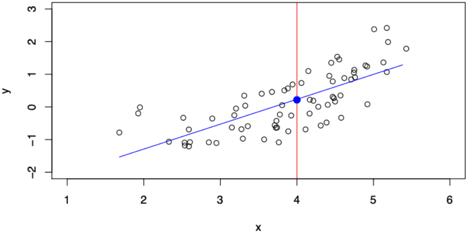
The linear model is an important example of a parametric model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

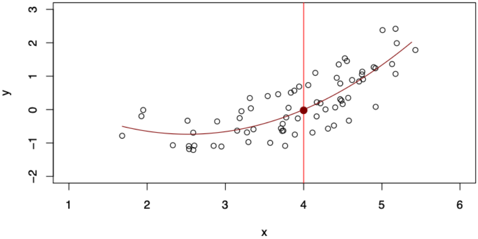
- First, a linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, \beta_2, ..., \beta_p$
- Second, we estimate the parameters by fitting the model to *training data*.

Parametric methods examples

A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better



Non-parametric methods

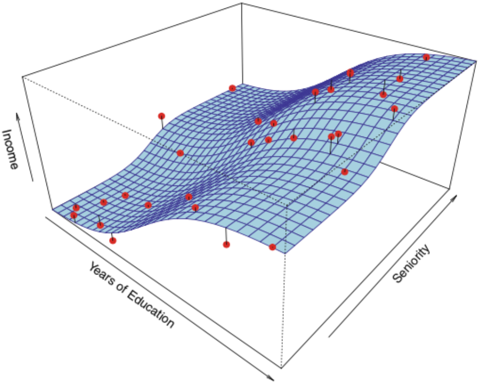
DO NOT make explicit assumptions about the functional form of f. Instead, seek an f that gets as close to the data points as possible without being too rough or wiggly.

Non-parametric methods

DO NOT make explicit assumptions about the functional form of f. Instead, seek an f that gets as close to the data points as possible without being too rough or wiggly.

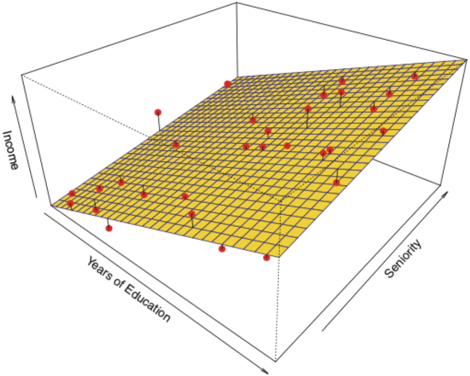
- Advantage: Avoiding the assumption of a particular functional form of f, have the potential to accurately fit a wider range of possible shapes for f.
- Parametric methods always bring the possibility that adopted f is very different from the true f.
- Disadvantage: use large nubmer of parameters, need large number of observations [*expensive*].

Example: finding a good f



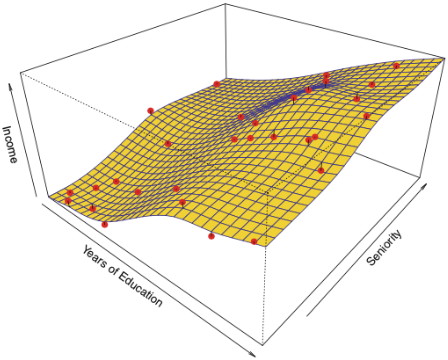
Income, years of education, and seniority

Example: finding a good f



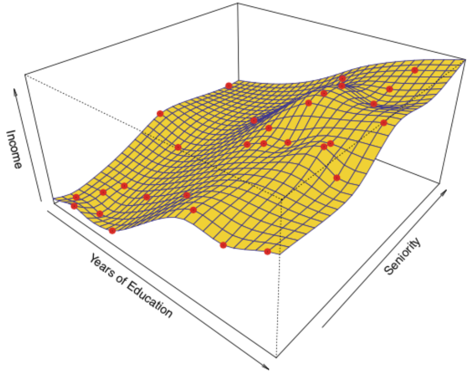
Linear fit

Example: finding a good f



A smooth thin-plate spline fit. Reasonable level of smoothness.

Example: finding a good f

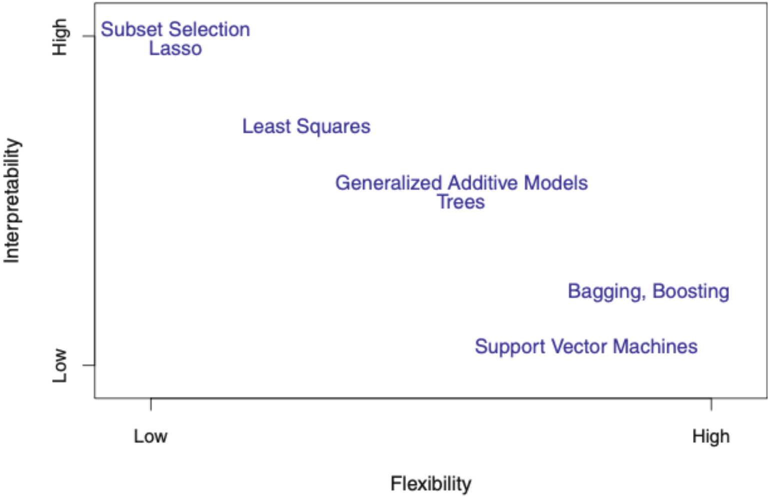


A smooth thin-plate spline fit. Low level of smoothness. Overfitting here!

Some trade-offs in modeling

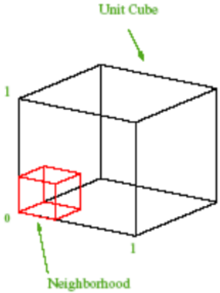
- Prediction accuracy versus interpretability.
  — Linear models are easy to interpret; thin-plate splines are not.
- Good fit versus over-fit or under-fit.
  — How do we know when the fit is just right?
- Parsimony versus black-box.
  — We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Flexibility vs. Interpretability



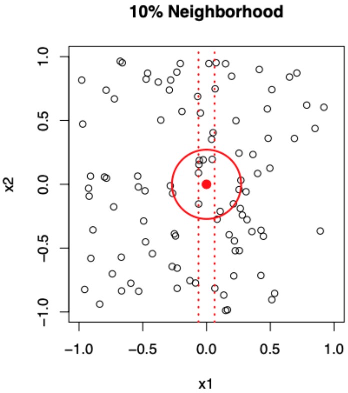
The curse of dimensionality

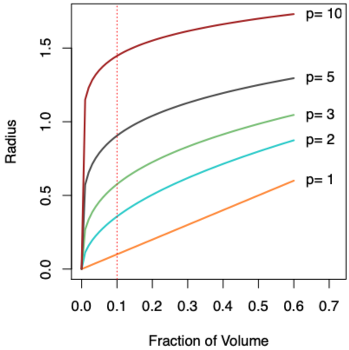
Richard Bellman, 1961



It is a luxury to have neighbors in higher dimensions!
In p-dimensions, to get a hypercube with volume r, the edge length needed is $r^{1/p}$.
In 10 dimensions, to capture 1% of the data to get a local average, we need 63% of the range of each input variable.

The curse of dimensionality





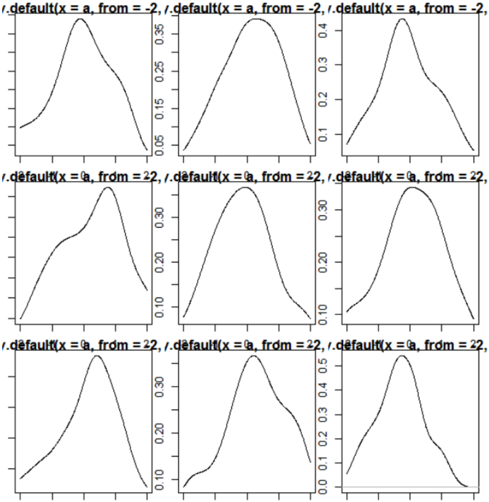
It is a luxury to have neighbors in higher dimensions!

The curse of dimensionality

In other words,
To get a “dense” sample, if we need $N = 100$ samples in 1 dimension, then we need $N = 100^{10}$ samples in 10 dimensions.
In high-dimension, the data is always sparse and do not support density estimation.
More data points are closer to the boundary, rather than to any other data point → prediction is much harder near the edge of the training sample.

The curse of dimensionality

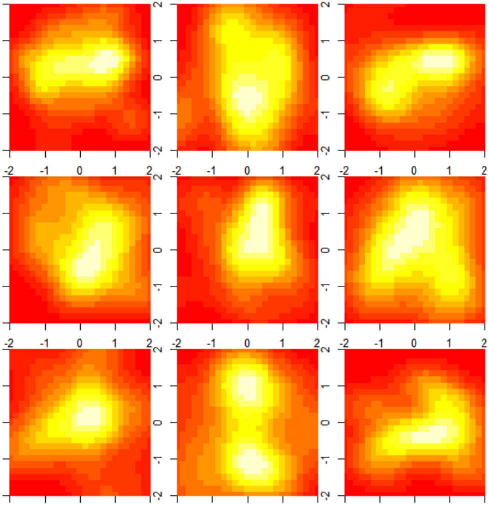
Estimating a 1D density with 40 data points.
Standard Normal distribution.



The curse of dimensionality

Estimating a 2D density with 40 data points.

2D normal distribution; zero mean; variance matrix is identity matrix.



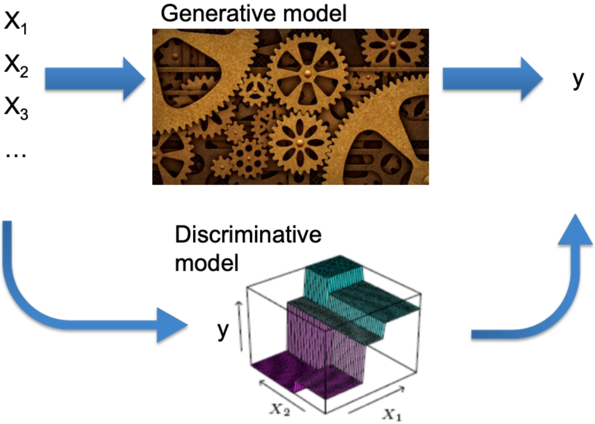
The curse of dimensionality

We have talked about the curse of dimensionality in the sense of density estimation.

In a classification problem, we do not necessarily need density estimation.

- Generative model — care about the mechanism: class density function.
  — Learns $p(\mathbf{X}, y)$, and predict using $p(y|\mathbf{X})$
  — In high dimensions, this is difficult.
- Discriminative model — care about boundary.
  — Learns $p(y|\mathbf{X})$ directly, potentially with a subset of $\mathbf{X}$

The curse of dimensionality



X_1

X_2

X_3

...

Generative model

y

Discriminative model

Example: Classifying sea bass and salmon. Looking at the length/width ratio is enough. Why should we care how many teeth each kind of fish have, or what shape fins they have?

Assessing Model Accurary

No one method dominates all others over all possible dataset [*No free lunch in statistics*].

Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.

Measuring the quality of fit

To evaluate how well our predictions actually match the observed data.
Mean Squared Error (MSE)

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

Training vs. Testing

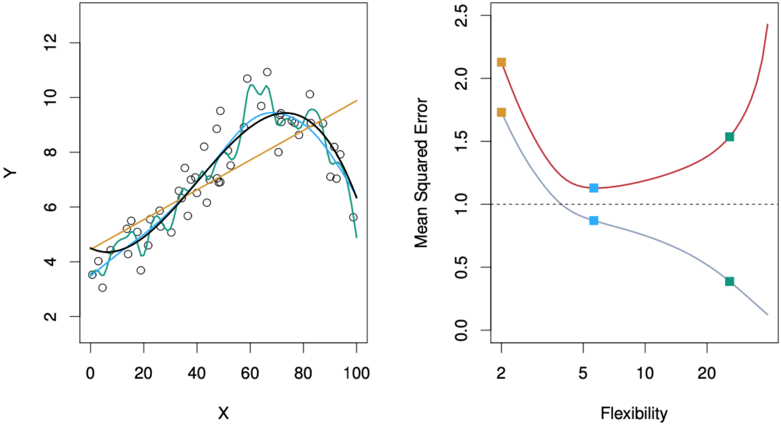
- We need to use some data to “train” our prediction model.
- Training: Estimating the parameters in the model.
- In general, we **DO NOT** care how well the method works in the training dataset.
- The accurary is more interesting when applying our model on **previously unseen dataset**.
- We need **both** training dataset and testing dataset.

Training vs. Testing examples

- Stock price: An algorithm to predict a stock's price based on previous (6 month) stock returns.
  — How well it predict last week's return ×
  — How well it predict tomorrow's return ✓
- Diabetes risk: An algorithm to predict whether a person's risk of having diabetes, using clinical measurements (e.g. weight, bp, age, family history, etc.).
  — How well it predict patients used to train this algorithm ×
  — How well it predict a **new person's** risk ✓

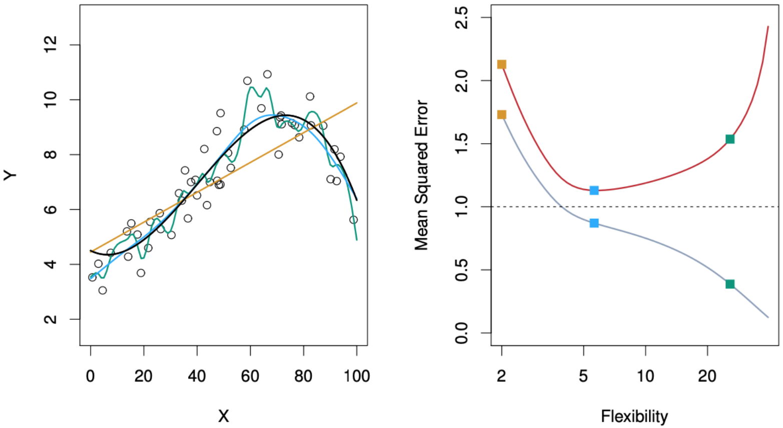
Method selection: minimize the testing MSE

Identify the method that minimize the **test** MSE.

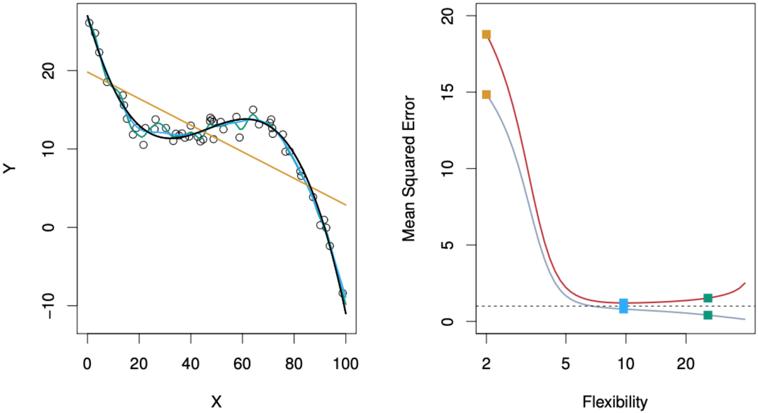


Black curve: truth. RHS: Red — Testing MSE. Grey — Training MSE.

Low training MSE does not guarantee low testing MSE



Training MSE keep decreasing while testing MSE ramps up.



Here the truth is wiggly and the noise is low, so the flexible fits do the best.

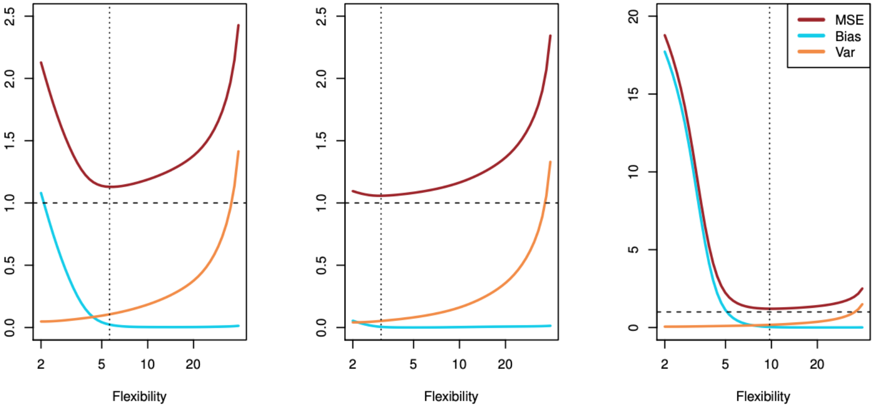
Bias-Variance Trade-off

The test MSE is the result of two competing properties in statistics: **Bias and Variance.**
Suppose we have a estimate $\hat{f}(x)$ for $f(x)$.
A new data point x_0 comes in.
The *expected test MSE* is then (show this!):

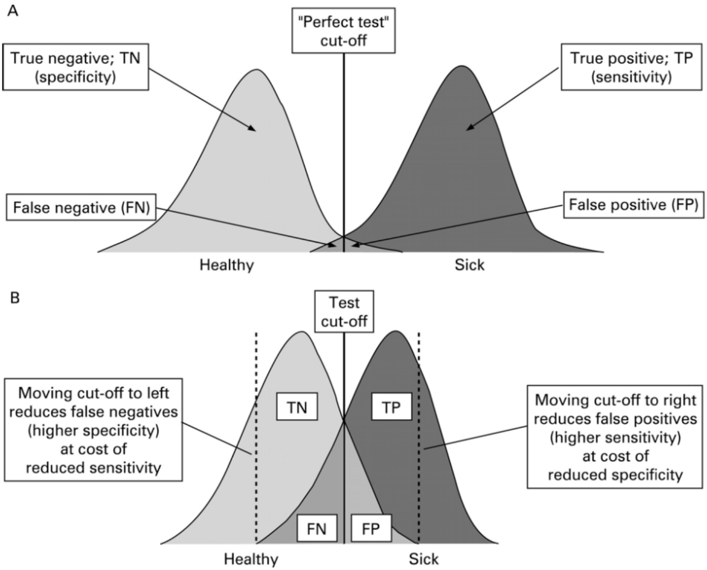
$$E(y_0 - \hat{f}(x_0))^2 = Var[\hat{f}(x_0)] + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

The 3 sources of test MSE: (1) sampling variation (variance) (2) choice during modeling (bias) and (3) irreducible error.
Given a fixed dataset, a more complex/flexible model tends to reduce bias but increase variation.

Bias-Variance Trade-off examples

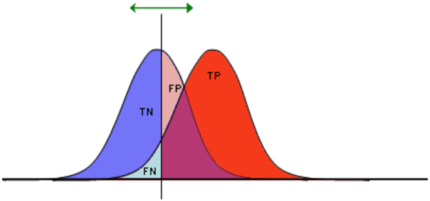


Reminder of the ROC curve

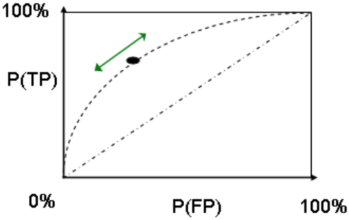

A
"Perfect test" cut-off
True negative; TN (specificity)
True positive; TP (sensitivity)
False negative (FN)
False positive (FP)
Healthy
Sick
B
Test cut-off
Moving cut-off to left reduces false negatives (higher specificity) at cost of reduced sensitivity
TN
TP
Moving cut-off to right reduces false positives (higher sensitivity) at cost of reduced specificity
FN
FP
Healthy
Sick

Reminder of the ROC curve

Receiver Operating Characteristic (ROC) curve



TP	FP
FN	TN
1	1



Textbook chapters

- ISLR: chapter 2: 2.1 - 2.2