# STA 4241 Lecture, Week 1

August 23rd, 2021

# Overview of what we will cover

- **Review of basic concepts and matrix computations**
- Introduction to statistical learning

# Review of key statistical concepts

- There are a number of concepts from probability theory that are used commonly throughout the class

- Random variables
  - Probability mass function / probability density function
  - Expected value
  - Variance
  - Rules for means and variances

- Conditional probability

# Statistics and estimators

- A statistic is simply a function of the observed data
    - statistics are random variables
- The standard deviation of a statistic is commonly referred to as its standard error
- Frequently in this class we will be using or finding estimates of unknown parameters $\theta$
- We will denote estimates of these unknown parameters as $\widehat{\theta}$
    - $\widehat{\theta}$ will be a function of the observed data

# Sampling distribution

- Suppose we estimate $\widehat{\theta}$ from a sample of $n$ observations
- The sampling distribution for $\widehat{\theta}$ is simply its probability distribution
- Characterizes the range of values we can expect to see for the statistic and how likely each of them are
- Imagine collecting a sample of $n$ observations a very large number of times, each time keeping track of $\widehat{\theta}$
    - The distribution of values you obtain is the sampling distribution

# Simple example

- Suppose we observe an independent sample of $n = 10$ observations from a normal distribution with variance 1 and unknown mean, $\theta$
- We are interested in using data to estimate $\theta$
- We will use the sample mean as our estimator

$$\widehat{\theta} = \frac{1}{10} \sum_{i=1}^{10} Y_i$$

# Mean and variance

- First, let's calculate the mean of our estimator

$$\begin{aligned}
E(\widehat{\theta}) &= E\left(\frac{1}{10}\sum_{i=1}^{10} Y_i\right) \\
&= \frac{1}{10}E\left(\sum_{i=1}^{10} Y_i\right) \\
&= \frac{1}{10}\sum_{i=1}^{10} E(Y_i) \\
&= \frac{1}{10}\sum_{i=1}^{10} \theta \\
&= \theta
\end{aligned}$$

- Our estimator is unbiased for $\theta$

# Mean and variance

- The variance follows a similar strategy

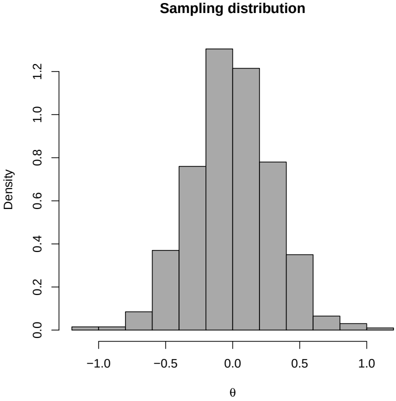
$$
\begin{aligned}
\mathrm{Var}(\widehat{\theta}) &= \mathrm{Var}\left(\frac{1}{10}\sum_{i=1}^{10} Y_i\right) \\
&= \frac{1}{100}\mathrm{Var}\left(\sum_{i=1}^{10} Y_i\right) \\
&= \frac{1}{100}\sum_{i=1}^{10} \mathrm{Var}(Y_i) \\
&= \frac{1}{100}\sum_{i=1}^{10} 1 \\
&= \frac{1}{10}
\end{aligned}
$$

# Mean and variance

- The mean of a sample of normal random variables is itself a normal random variable
- Therefore the sampling distribution of $\widehat{\theta}$ is normal with mean $\theta$ and variance $1/10$
- To see this empirically, let's simulate 1000 samples of size $n = 10$ from a normal distribution with mean 0 and variance 1
  - See weekly R code for how to do this

# Sampling distribution

- The mean of our samples is 0.0137 and the variance is 0.099888
  - Very close to the true values
- Below is a histogram of the 1000 $\widehat{\theta}$ values obtained



# Linear algebra

- We will denote matrices with bold letters
- Here, $\boldsymbol{X}$ is an $n \times p$ matrix

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

- $X_{ij}$ is the element of $\boldsymbol{X}$ corresponding to the ith row and jth column

# Linear algebra

- We will commonly use the transpose of a matrix

$$\boldsymbol{X}^T = \begin{pmatrix} X_{11} & X_{21} & \dots & X_{n1} \\ X_{12} & X_{22} & \dots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \dots & X_{np} \end{pmatrix}$$

- The transpose of an $n \times p$ matrix $\boldsymbol{X}$ is a $p \times n$ matrix that has the rows and columns switched
    - The rows of $\boldsymbol{X}$ are the columns of $\boldsymbol{X}^T$

# Linear algebra

- Matrix multiplication will also be used in the course regularly
- If we want to multiply an $n \times p$ matrix $\boldsymbol{X}$ and a $p \times k$ matrix $\boldsymbol{Y}$, then the $(i, j)$ element is defined by $\sum_{k=1}^{p} X_{ik} Y_{kj}$
- The number of columns in $\boldsymbol{X}$ has to be the same as the number of rows in $\boldsymbol{Y}$

$$\begin{aligned}\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} &= \begin{pmatrix} 1 \times 5 + 2 \times 7 & 1 \times 6 + 2 \times 8 \\ 3 \times 5 + 4 \times 7 & 3 \times 6 + 4 \times 8 \end{pmatrix} \\ &= \begin{pmatrix} 19 & 22 \\ 43 & 50 \end{pmatrix}\end{aligned}$$

# Linear algebra

- An inverse of any square $n \times n$ matrix is a matrix $\boldsymbol{A}^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n$
- There are extensions of inverses to non-square matrices, but we will only restrict attention to square matrices
- Not all matrices have inverses
  - These are called singular matrices
  - Generally won't encounter this often in class

# Linear algebra

- We also denote vectors with bold notation
- A very common one we will encounter is

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

- $\boldsymbol{Y}$ will frequently denote our outcome of interest, i.e. the dependent variable

# Overview of what we will cover

- Review of probability and matrix computations
- **Introduction to statistical learning**

# What is statistical learning?

- This class is about learning a large set of tools to help analyze and model data

- We will learn not only how to use these tools, but how to choose which tool is appropriate for a given situation

- Much of the focus in this class will be on prediction
    - Given some inputs, what do I expect the output to be

- There are many other goals in statistics, some of which we will cover in this class

# What is statistical learning?

- A large portion of this class will center on the following model

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

- $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ are $p$ distinct predictors
- $f$ is an unknown function and is generally the quantity of interest
  - Given some predictors, what do I expect the outcome to be?
- $\epsilon$ is a vector of error terms that have mean zero and are independent of $\boldsymbol{X}$
- We will use observed data to estimate $f$

# What is statistical learning?

- This may seem straightforward but there are many issues that need to be dealt with or are themselves of scientific interest
- Which predictors in $\boldsymbol{X}$ are most useful for predicting $Y$?
- Do I need to use nonlinear models, or will a linear approximation to $f$ work well
- Do I care about the interpretation of $f$, or only how well it predicts $Y$?

# What is our goal?

- When we fit this model, we need to make sure we understand why we are interested in estimating $f$
- There are two main reasons to estimate $f$
  - Prediction - using inputs to predict outputs
  - Inference - understanding relationships between variables
- These goals are very different
  - One estimate of $f$ may be great at prediction but provide very little information about relationships between variables

# Prediction

- Suppose we want to predict the outcome, so we estimate $\widehat{f}$
- Given a set of inputs, $X$, we can estimate the outcome via

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

- If we're interested in prediction, the only thing we care about is the magnitude of $Y - \widehat{Y}$
- Larger absolute values imply worse predictions/model fits

# Prediction

- One metric commonly looked at is the average squared difference
  - Mean squared error (MSE)

- Assume for now that $\boldsymbol{X}$ and $\widehat{f}$ are fixed

$$\begin{aligned}
E[(Y - \widehat{Y})^2] &= E[(f(\boldsymbol{X}) + \epsilon - \widehat{f}(\boldsymbol{X}))^2] \\
&= E[(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}) + \epsilon)^2] \\
&= E[(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + \epsilon^2 + 2\epsilon(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))] \\
&= E[(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2] + E[\epsilon^2] + 2E[\epsilon(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))] \\
&= (f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + \text{Var}(\epsilon) + 2E[\epsilon]E[f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X})] \\
&= (f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + \text{Var}(\epsilon)
\end{aligned}$$

# Prediction

- This error breaks into two parts

$$(f(\boldsymbol{X}) - \widehat{f}(\boldsymbol{X}))^2 + \text{Var}(\epsilon)$$

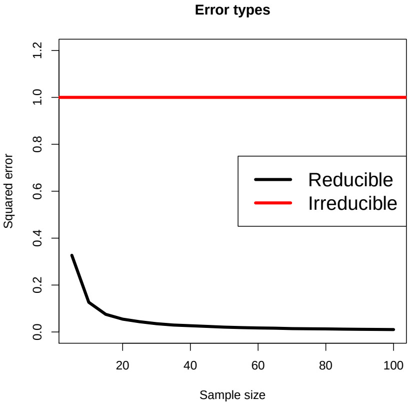
- We have no control over $\epsilon$
  - Called the irreducible error or the noise
  - Contains unmeasurable variation, or the effects of unmeasured predictors

- The first source of error, however, can be reduced

- Our goal is to minimize the first source of error
  - Sometimes called the reducible error or the signal

# Prediction

- This figure highlights both sources of error from a simulated example
- The reducible error goes to zero as the sample size increases
- Regardless of sample size, the irreducible error will not go away



# Prediction or inference?

- If our goal is simply predicting the outcome, then we do not care about the form of $f$, we only care how well it minimizes the reducible error
  - $f$ is a black box
  - No interpretability
  - Many complex machine learning prediction algorithms are like this
- Many times, we are interested in what $f$ looks like
  - Is the relationship linear?
  - What is the relationship between $X_j$ and $Y$?
  - Is this relationship constant at different levels of the other covariates?
  - Are all predictors important, or only some of them?

# Prediction or inference?

- Not all problems fall squarely into one or the other
- Many times we want to make good predictions but also learn about the structure of the $f$ function
- Suppose we have predictors $\boldsymbol{X}$ that are environmental risk factors and our outcome $Y$ is whether or not a patient has a particular illness
  - Clearly, we are interested in prediction so we can predict who is most likely to get sick
  - Also interested in the manner in which the risk factors affect illness rates so we can intervene to reduce future illness

# Estimating $f$

- Whether we are interested in prediction, inference, or both heavily drives our decision on how we estimate $f$
- There are a wide range of methods to estimate $f$
    - Most of this course is dedicated to it!
- Each approach has varying statistical properties that may be useful to us in any particular situation
    - How do we choose which one to use?

# Estimating $f$

- There is generally a trade-off between model interpretability and prediction accuracy
- Simpler models might be more interpretable and thus more amenable for doing inference
- Complex models tend to function more like a black box and provide very good predictions, but very little in terms of understanding the $f$ function
- If we are solely interested in inference, we might prefer the simpler models
    - Not universally true, but a good general rule

# Parametric models

- Parametric methods make assumptions about the functional form of $f$

- The most common assumption is that of linearity

$$f(\boldsymbol{X}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

- This greatly simplifies the problem of estimating $f$
  - Only need to estimate $p+1$ parameters instead of an arbitrary function
    - This is why we call them parametric

# Parametric models

- This simplification comes at a cost
- What if the true $f$ is far from linear?
  - $\widehat{f} - f$ will be large
  - Large reducible error and poor predictions
- What if the true $f$ is well approximated by a linear model?
  - More efficient estimates than complex models
  - Easily interpretable parameters
- Next week we will review the linear model in detail

# Nonparametric models

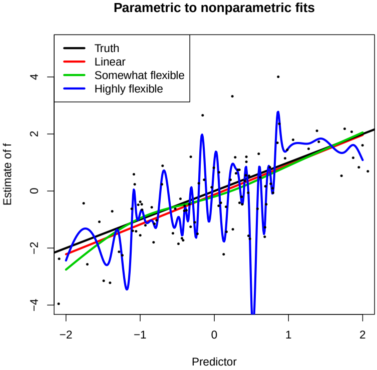
- We also have nonparametric models, which do not make parametric assumptions about $f$
- This avoids the issue of misspecifying the functional form of $f$
    - Should lead to good predictions
- Nonparametric models aim to produce fits that are close to the observed data, without overfitting

# Nonparametric models

- Nonparametric approaches frequently make some assumptions about $f$, though they are much weaker than those from parametric models
  - Smoothness, Shape constraints, others
- The main issue with nonparametric models is to avoid overfitting
- Overfitting occurs when our estimated $f$ is fit too closely to the observed data points
  - Leads to bad out-of-sample predictions
- Nonparametric models typically require bigger sample sizes than parametric models

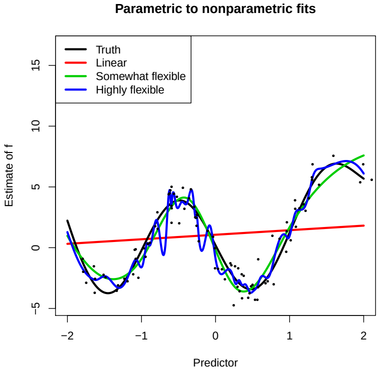
# Example highlighting overfitting

- Suppose we have one predictor $X_1$ and the true $f$ is linear
- We try fitting three models that range from linear to highly nonlinear
- The highly nonlinear model is substantially overfit to the observed data points (blue curve)



# Example highlighting when linearity fails

- Now the true $f$ is very nonlinear
- Linear model does extremely poorly
- Green line corresponding to a moderate degree of nonlinearity fits the data quite well
  - Blue line still somewhat overfit

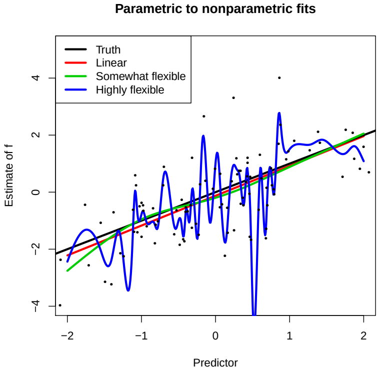


# How to choose an approach?

- The previous two examples show us that it is crucially important that we choose the right degree of flexibility for our model
- How do we go about choosing which method is best?
  - We won't know the true curve in practice
- Suppose we want our approach to minimize the MSE
  - Minimize $E[(Y - \widehat{f}(\boldsymbol{X}))^2]$
  - In practice, we minimize the sample version, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{f}(\boldsymbol{X}_i))^2$
- This seems reasonable because we want to estimate $f$ as well as possible

# Looking at MSE

- Let's return to the linear example
- The in-sample MSE of the linear model is 1.12 while it is only 0.60 for the highly flexible model!



# Out of sample MSE is much more important

- We can tell visually that the linear fit is much closer to the truth than the highly flexible one
- The in-sample MSE of the flexible model is low because it is overfit to the observed data
    - We could make it zero if we used a flexible enough model!
- The out of sample MSE is a more relevant quantity to examine
    - $\boldsymbol{X}_0$ is a new set of predictors and $Y_0$ is the corresponding outcome
    - Want to minimize $E[(Y_0 - \widehat{f}(\boldsymbol{X}_0))^2]$
- The out of sample MSE is 1.77 for the flexible model and 1.1 for the linear model

# Out of sample MSE

- The out of sample MSE can be decomposed as

$$E[(Y_0 - \widehat{f}(\boldsymbol{X}_0))^2] = \text{Var}(\widehat{f}(\boldsymbol{X}_0)) + [\text{Bias}(\widehat{f}(\boldsymbol{X}_0))]^2 + \text{Var}(\epsilon)$$

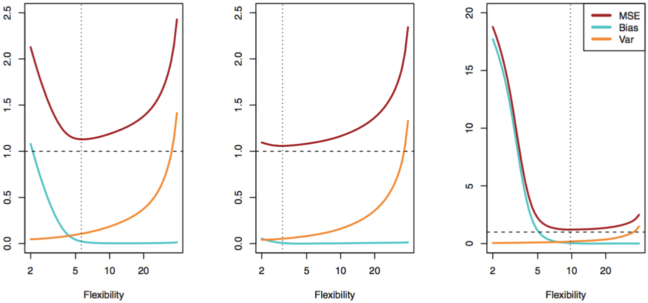
- We have no control over random error $\epsilon$
  - Regardless of sample size or model choice
- We do have control over bias and variance of our estimators
- Ideally we want an estimator with low bias and low variance

# Bias-variance trade-off

- There is a bias-variance trade-off that exists with most statistical approaches
- Simpler approaches usually have small variance, but may be biased
- Complex approaches are less susceptible to bias, but are usually less efficient (higher variance)
- Our goal will be to find an estimator that balances these competing concerns in an optimal way

# Bias-variance trade-off

- Here is a figure from the textbook showing this trade-off explicitly for three different data sets
- Variance increases with flexibility, but bias decreases
- MSE is optimized at various levels of flexibility for the three examples



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

# Classification

- Nearly all of the previous discussion was centered around continuous outcomes
    - Sometimes called the regression setting
- Another very important topic centers around categorical outcomes
    - Extremely common scenario in practice
- In this scenario, we are interested in using inputs to predict class membership
    - Referred to as classification

# Classification

- We still want to estimate a function $f$ that helps us predict the outcome, i.e. class membership

- One way to assess the accuracy of any approach is the training classification error rate

$$\frac{1}{n}\sum_{i=1}^{n} 1(Y_i \neq \widehat{Y}_i)$$

- $1(Y_i \neq \widehat{Y}_i)$ is an indicator variable for whether our prediction does not equal the true value
    - $1(Y_i \neq \widehat{Y}_i) = 1$ if $Y_i \neq \widehat{Y}_i$
    - $1(Y_i \neq \widehat{Y}_i) = 0$ if $Y_i = \widehat{Y}_i$

# Classification

- Similar to the continuous outcome setting, the training error rate is susceptible to overfitting and is typically not of interest to us
- The test error rate is the quantity we actually want to minimize

$$E[1(Y_0 \neq \widehat{Y}_0)]$$

- Where $Y_0$ represents an out of sample data point
- A good classification algorithm makes this quantity small

# Bayes Classifier

- Suppose we knew the following probability

$$P(Y = j|X = x_0)$$

- We could then set our prediction to be the value of $j$ that maximizes this expression
  - Set $\widehat{Y}$ to its most likely value

- This is called the Bayes classifier

# Bayes Classifier

- The Bayes classifier is the best we could possibly hope to achieve
- For a particular $x_0$ value, the Bayes error rate is $1 - \text{max}_j P(Y = j|X = x_0)$
- Think about a case where $Y$ only takes values 0 or 1 and $P(Y = 1|X = x_0) = 0.7$
  - Our prediction is $\widehat{Y} = 1$
  - This will be wrong in the 30% of the time that $Y = 0$
  - Therefore the error rate is $1 - 0.7 = 0.3$

# Bayes Classifier

- The average error rate is then simply

$$1 - E(\text{max}_j P(Y = j|X))$$

- The expectation is with respect to the distribution of $X$
- This average error rate is the classification equivalent of $\text{Var}(\epsilon)$ from the continuous setting
- This is the irreducible error that our model can't hope to lower

# Bayes Classifier

- The Bayes classifier is not feasible in practice
  - We never know the true conditional probability
- However, we can estimate this conditional probability
- Once we have an estimate of the probability, we can proceed in the same manner
- The K-nearest neighbors classifier is a relatively simple approach that estimates $P(Y = j|X)$

# K-nearest neighbors (KNN)

- Suppose we choose a positive integer $K$
- Our interest is classifying a subject with $X = x_0$
- KNN finds the $K$ subjects in the data whose $X$ values are closest to $x_0$
  - Denote these by $\mathcal{N}_0$
- KNN estimates the conditional probability as

$$P(Y \widehat{= j | X} = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} 1(Y_i = j)$$

# K-nearest neighbors (KNN)

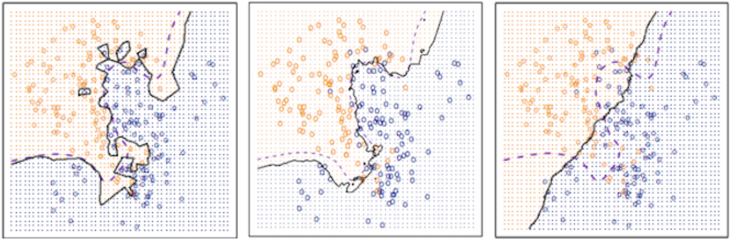
- By finding the $K$ closest subjects with respect to $X$, we are effectively conditioning on $X$
- Then, once we have effectively conditioned on $X$, we simply take the observed proportion of data that has $Y_i = j$ as our estimate
- There are two sources of error for our estimate of $P(Y = j|X = x_0)$
  - The $K$ subjects don't have exactly $X_i = x_0$ (bias)
  - Taking a sample average as an estimate of the true proportion (variance)

# K-nearest neighbors (KNN)

- The magnitude of these two sources of error is dictated by $K$
- Larger $K$ means less variability in taking a sample proportion
- Larger $K$ also means we have to use subjects whose $X$ values are farther from $x_0$
- Bias-variance trade-off!
  - Dictated by $K$
- The book has a great example illustrating this bias-variance trade-off

# K-nearest neighbors (KNN)

- There are two covariates $X_1$ and $X_2$ represented by the two axes
- The purple line is the Bayes classifier boundary while the black lines are corresponding KNN estimates for $K = 1, 10, 100$
  - The boundary line is the line where the classifier switches from predicting orange to blue



James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

# K-nearest neighbors (KNN)

- $K = 10$ does a very good job of approximating the ideal Bayes classifier
- Setting $K$ too small makes the boundary overly flexible
  - Overfit to the observed data
  - High variance
  - Extremely small training error rate
- Setting $K$ too large makes the boundary too smooth
  - Low variance, but higher bias

# How do we decide which approach to use

- All of the methods we have shown have decisions that need to be made
    - Degree of flexibility in estimating $f$
    - Choosing $K$ for KNN
- The optimal choice depends highly on unknown features of the data
- For now, it suffices to understand that these choices are important and that there are trade-offs occurring when we choose these parameters
- In a few weeks, we will learn ways to make these decisions using the observed data

# Supervised vs. unsupervised learning

- Everything that we have talked about so far falls in the category of supervised learning
  - Utilized the outcome
- In some cases, the outcome is not measured
  - Unsupervised settings
- There are fewer ways to analyze data in unsupervised settings
  - Clustering, principle components analysis
- We will mostly focus on supervised settings in this class, though we will briefly touch on unsupervised settings as well