Supervised Learning: Introduction

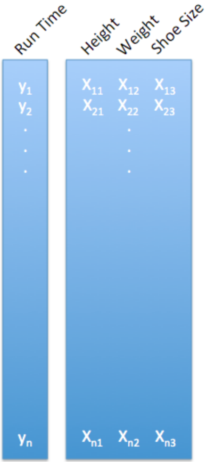
Noah Simon & Ali Shojaie

July 17-19, 2019
Summer Institute for Statistics of Big Data
University of Washington

A Simple Example

- Suppose we have $n = 500$ kids for whom we have $p = 3$ measurements: height, weight, and shoe size.
- We wish to predict these kids' 1600-meter run times using these measurements.

A Simple Example



Notation:

- n is the number of observations.
- p the number of variables/features/predictors.
- y is a n-vector containing response/outcome for each of n observations.
- X is a $n \times p$ data matrix.

Linear Regression on a Simple Example

- You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

  where y is run time, X_1, X_2, X_3 are height, weight, and shoe size, and ϵ is a noise term.

Linear Regression on a Simple Example

- You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

  where y is run time, X_1, X_2, X_3 are height, weight, and shoe size, and ϵ is a noise term.
- You can look at the coefficients, p-values, and t-statistics for your linear regression model in order to interpret your results.

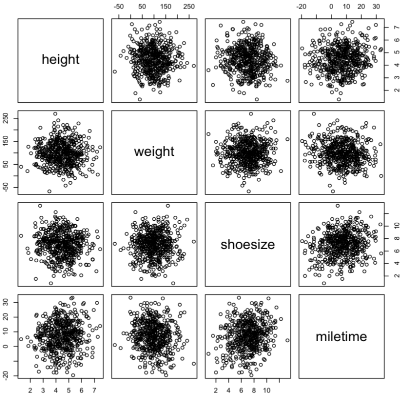
Linear Regression on a Simple Example

- You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

  where y is run time, X_1, X_2, X_3 are height, weight, and shoe size, and ϵ is a noise term.
- You can look at the coefficients, p-values, and t-statistics for your linear regression model in order to interpret your results.
- You learned everything (or most of what) you need to analyze this data set in AP Statistics!

A Relationship Between the Variables?



Linear Model Output

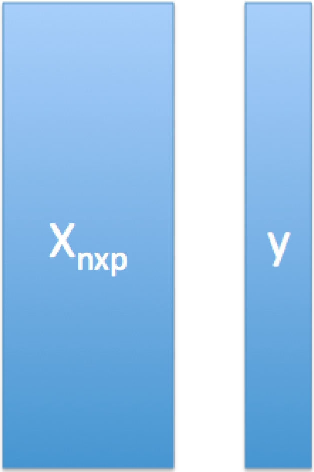
	Estimate	Std. Error	T-Stat	P-Value
Intercept	-2.265831	2.644654	-0.857	0.39199
height	1.074814	0.414789	2.591	0.00985 **
weight	-0.021155	0.008482	-2.494	0.01295 *
shoesize	0.955222	0.214449	4.454	1.04e-05 ***

$$\mathrm{RunTime} \approx -2.27 + 1.07 \times \mathrm{Height} - 0.021 \times \mathrm{Weight} + 0.96 \times \mathrm{ShoeSize}.$$

Low-Dimensional Versus High-Dimensional

- The data set that we just saw is low-dimensional: $n \gg p$.
- Lots of the data sets coming out of modern biological techniques are high-dimensional: $n \approx p$ or $n \ll p$.
- This poses statistical challenges! AP Statistics no longer applies.

Low Dimensional



High Dimensional

What Goes Wrong in High Dimensions?

- Suppose that we included many additional predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2

What Goes Wrong in High Dimensions?

- Suppose that we included many additional predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2
- Some of these predictors are useful, others aren't.

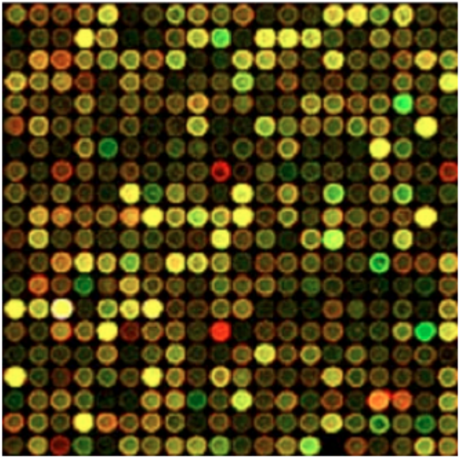
What Goes Wrong in High Dimensions?

- Suppose that we included many additional predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2
- Some of these predictors are useful, others aren't.
- If we include too many predictors, we will overfit the data.
- Overfitting: Model looks great on the data used to develop it, but will perform very poorly on future observations.

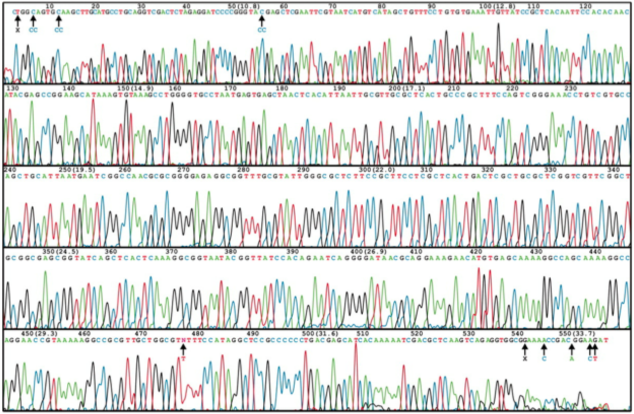
What Goes Wrong in High Dimensions?

- Suppose that we included many additional predictors in our model, such as
  - 50-yard dash time
  - Age
  - Zodiac symbol
  - Favorite color
  - Mother's birthday, in base 2
- Some of these predictors are useful, others aren't.
- If we include too many predictors, we will overfit the data.
- Overfitting: Model looks great on the data used to develop it, but will perform very poorly on future observations.
- When $p \approx n$ or $p > n$, overfitting is guaranteed unless we are very careful.

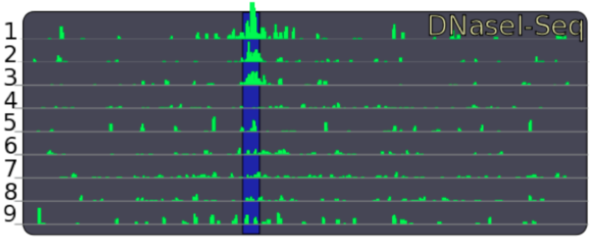
Gene Expression Data



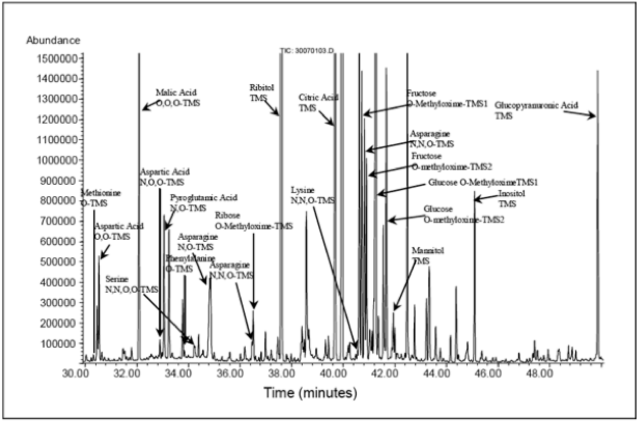
DNA Sequence Data



DNAse Hypersensitivity Data



Metabolomic Data



High-Dimensional Omics Analyses

For most omics analyses, we have many more variables than observations.... i.e. $p \gg n$.

High-Dimensional Omics Analyses

For most omics analyses, we have many more variables than observations.... i.e. $p \gg n$.

- Predict risk of diabetes on the basis of DNA sequence data.... using $n = 1000$ patients and $p = 3000000$ variables.

High-Dimensional Omics Analyses

For most omics analyses, we have many more variables than observations.... i.e. $p \gg n$.

- Predict risk of diabetes on the basis of DNA sequence data.... using $n = 1000$ patients and $p = 3000000$ variables.
- Cluster tissue samples on the basis of DNase hypersensitivity... using $n = 200$ cell types and $p = 1000000000$ variables.

High-Dimensional Omics Analyses

For most omics analyses, we have many more variables than observations.... i.e. $p \gg n$.

- Predict risk of diabetes on the basis of DNA sequence data.... using $n = 1000$ patients and $p = 3000000$ variables.
- Cluster tissue samples on the basis of DNase hypersensitivity... using $n = 200$ cell types and $p = 1000000000$ variables.
- Identify genes whose expression is associated with survival time... using $n = 250$ cancer patients and $p = 20000$ variables.

Why Does Dimensionality Matter?

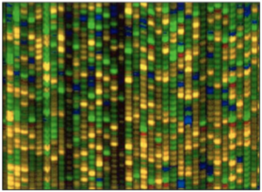
- Classical statistical techniques, such as linear regression, *cannot* be applied.
- Even very simple tasks, like identifying variables that are associated with a response, must be done with care.
- High risks of overfitting, false positives, and more.

Why Does Dimensionality Matter?

- Classical statistical techniques, such as linear regression, *cannot* be applied.
- Even very simple tasks, like identifying variables that are associated with a response, must be done with care.
- High risks of overfitting, false positives, and more.

This course: Statistical machine learning tools for big – mostly high-dimensional – data.

Statistical Machine Learning



Supervised and Unsupervised Learning

- Statistical machine learning can be divided into two main areas: supervised and unsupervised.

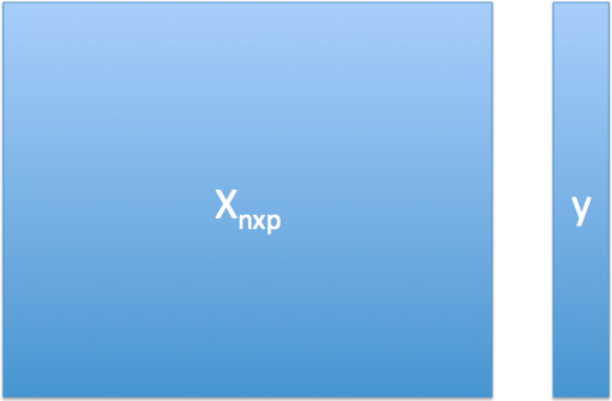
Supervised and Unsupervised Learning

- Statistical machine learning can be divided into two main areas: supervised and unsupervised.
- Supervised Learning: Use a data set X to predict or detect association with a response y.
  - Regression
  - Classification
  - Hypothesis Testing

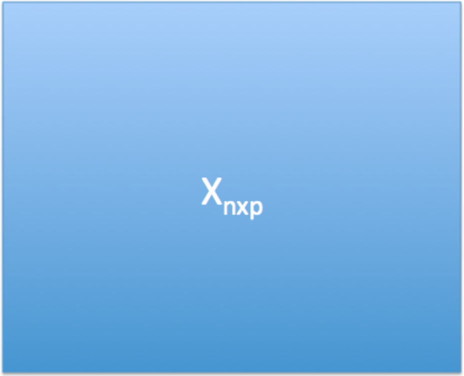
Supervised and Unsupervised Learning

- Statistical machine learning can be divided into two main areas: supervised and unsupervised.
- Supervised Learning: Use a data set X to predict or detect association with a response y.
  - Regression
  - Classification
  - Hypothesis Testing
- Unsupervised Learning: Discover the signal in X, or detect associations within X.
  - Dimension Reduction
  - Clustering

Supervised Learning



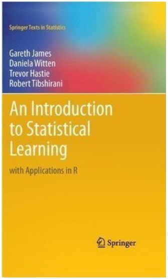
Unsupervised Learning



This Course

- We will cover the big ideas in supervised learning for big data.
- The best way to use these methods: learn R.

“Course Textbook” . . . with applications in R



- Available for (free!) download from www.statlearning.com.
- An accessible introduction to statistical machine learning, with an R lab at the end of each chapter!!
- We will go through some of these R labs in class.
- To learn more, go through them on your own!

Let's Try Out Some R!

Chapter 2 R lab
www.statlearning.com