

Supervised Learning: Introduction

Jean Feng & Ali Shojaie

Aug 19-21, 2024
Summer Institute in Statistics for Big Data
University of Washington

Slides

Slides and (some) codes are available at

<https://github.com/SISBID/Module4>

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

Run Time	Height	Weight	Shoe Size
y_1	x_{11}	x_{12}	x_{13}
y_2	x_{21}	x_{22}	x_{23}
.	.	.	.
.	.	.	.
.	.	.	.
y_n	x_{n1}	x_{n2}	x_{n3}

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

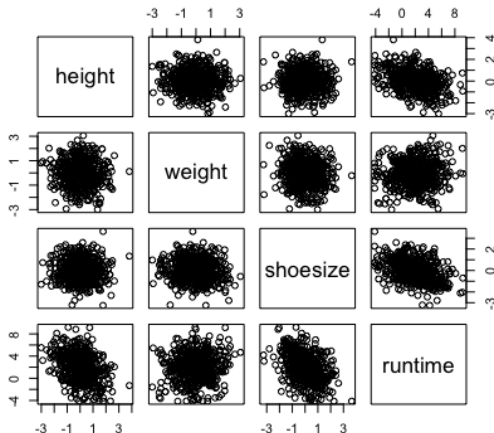
- ▶ You can perform linear regression to develop a model to predict run time using height, weight, and shoe size:

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

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	1.94179	0.09590	20.247	<2e-16 ***
height	-0.87704	0.09489	-9.243	<2e-16 ***
weight	0.07961	0.09105	0.874	0.382
shoesize	-1.00405	0.09530	-10.535	<2e-16 ***

$\text{RunTime} \approx 1.94 - 0.88 \times \text{Height} + 0.08 \times \text{Weight} - 1.00 \times \text{ShoeSize}.$

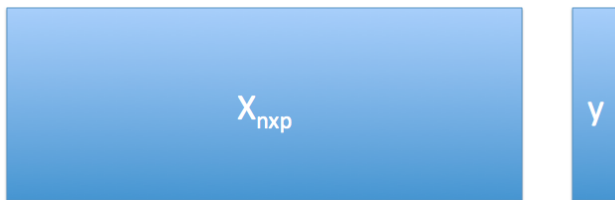
Low-Dimensional Versus High-Dimensional

- ▶ 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.

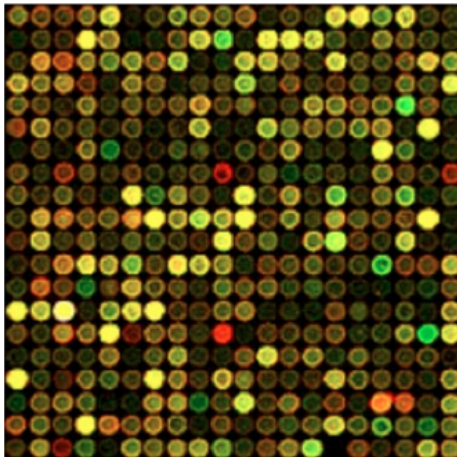
What Goes Wrong in High Dimensions?

- ▶ 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.

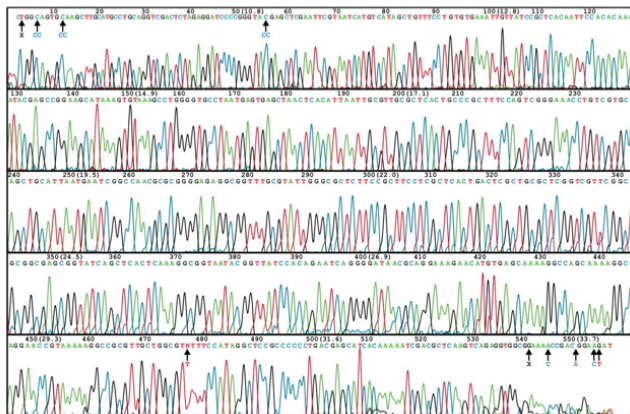
What Goes Wrong in High Dimensions?

- ▶ 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.

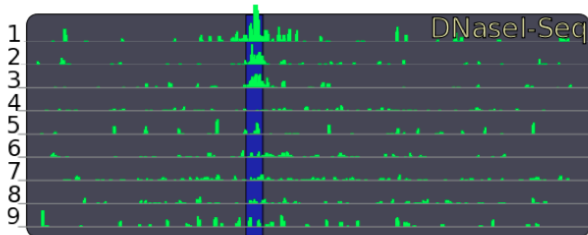
Gene Expression Data



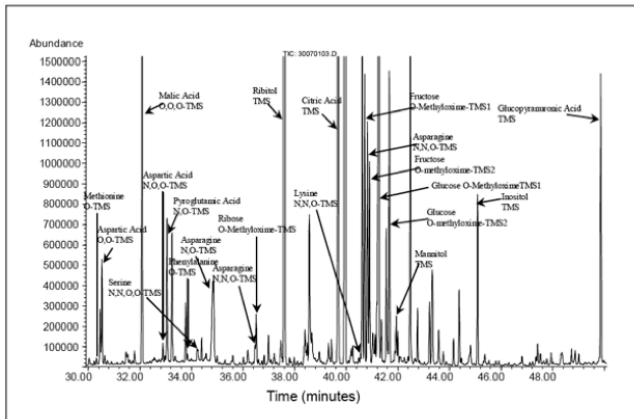
DNA Sequence Data



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?

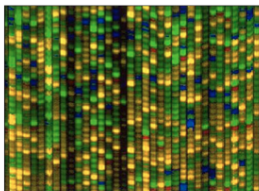
- ▶ Classical statistical techniques, such as linear regression, *cannot* be applied.
- ▶ Even very simple tasks, like identifying variables that are associated with a response, must be done with care.
- ▶ High risks of **overfitting**, **false positives**, and more.

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

Statistical Machine Learning



Google™



Supervised and Unsupervised Learning

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

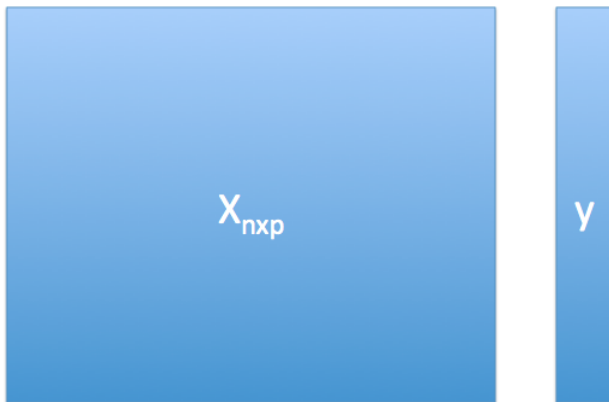
Supervised and Unsupervised Learning

- ▶ **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

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
- ▶ **Unsupervised Learning:** Discover the signal in X , or detect associations within X .
 - ▶ Dimension Reduction
 - ▶ Clustering
 - ▶ Hypothesis Testing

Supervised Learning



Unsupervised Learning



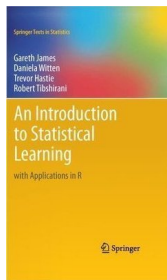
$X_{n \times p}$

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

Two Key Tasks in Supervised Learning: Regression & Classification

Two Key Tasks in Supervised Learning: Regression & Classification

- ▶ **Regression:** Predict a **quantitative** response, such as
 - ▶ blood pressure
 - ▶ cholesterol level
 - ▶ tumor size

Two Key Tasks in Supervised Learning: Regression & Classification

- ▶ **Regression:** Predict a **quantitative** response, such as
 - ▶ blood pressure
 - ▶ cholesterol level
 - ▶ tumor size
- ▶ **Classification:** Predict a **categorical** response, such as
 - ▶ tumor versus normal tissue
 - ▶ heart disease versus no heart disease
 - ▶ subtype of glioblastoma

Two Key Tasks in Supervised Learning: Regression & Classification

- ▶ **Regression:** Predict a **quantitative** response, such as
 - ▶ blood pressure
 - ▶ cholesterol level
 - ▶ tumor size
- ▶ **Classification:** Predict a **categorical** response, such as
 - ▶ tumor versus normal tissue
 - ▶ heart disease versus no heart disease
 - ▶ subtype of glioblastoma
- ▶ We will start with **Regression**.

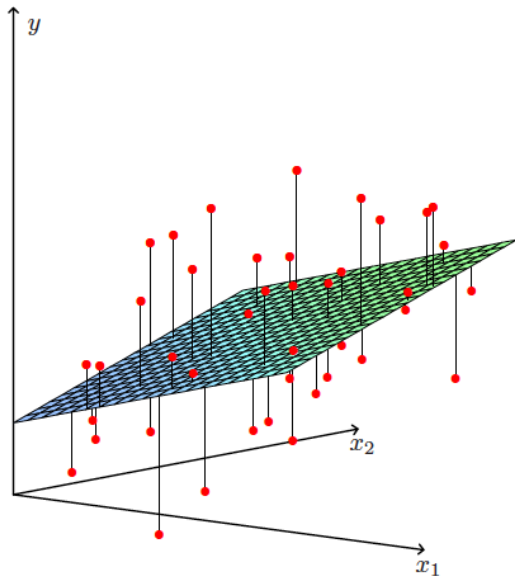
Linear Models

- ▶ We have n observations, for each of which we have p predictor measurements and a response measurement.
- ▶ Want to develop a model of the form

$$y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

- ▶ Here ϵ_i is a noise term associated with the i th observation.
- ▶ Must estimate $\beta_0, \beta_1, \dots, \beta_p$ — i.e. we must **fit the model**.

Linear Model With $p = 2$ Predictors



What Makes a Model Linear?

What Makes a Model Linear?

- ▶ A linear model is **linear in the regression coefficients!**

What Makes a Model Linear?

- ▶ A linear model is **linear in the regression coefficients!**
- ▶ This is a linear model:

$$y_i = \beta_1 \sin(X_{i1}) + \beta_2 X_{i2} X_{i3} + \epsilon_i.$$

What Makes a Model Linear?

- ▶ A linear model is **linear in the regression coefficients!**
- ▶ This is a linear model:

$$y_i = \beta_1 \sin(X_{i1}) + \beta_2 X_{i2} X_{i3} + \epsilon_i.$$

- ▶ This is not a linear model:

$$y_i = \beta_1^{X_{i1}} + \sin(\beta_2 X_{i2}) + \epsilon_i.$$

Linear Models in Matrix Form

- ▶ For simplicity, ignore the intercept β_0 .
 - ▶ Assume $\sum_{i=1}^n y_i = \sum_{i=1}^n X_{ij} = 0$; in this case, $\beta_0 = 0$.
 - ▶ Alternatively, let the first column of \mathbf{X} be a column of 1's.

Linear Models in Matrix Form

- ▶ For simplicity, ignore the intercept β_0 .
 - ▶ Assume $\sum_{i=1}^n y_i = \sum_{i=1}^n X_{ij} = 0$; in this case, $\beta_0 = 0$.
 - ▶ Alternatively, let the first column of \mathbf{X} be a column of 1's.
- ▶ In matrix form, we can write the linear model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

i.e.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Least Squares Regression

- ▶ There are a lot of ways we could fit the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- ▶ Most common approach in classical statistics is **least squares**:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right\}.$$

Here $\|\mathbf{a}\|^2 \equiv \sum_{i=1}^n a_i^2$.

Least Squares Regression

- ▶ There are a lot of ways we could fit the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$.
- ▶ Most common approach in classical statistics is **least squares**:

$$\underset{\beta}{\text{minimize}} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|^2 \right\}.$$

Here $\|\mathbf{a}\|^2 \equiv \sum_{i=1}^n a_i^2$.

- ▶ We are looking for β_1, \dots, β_p such that

$$\sum_{i=1}^n (y_i - (\beta_1 X_{i1} + \dots + \beta_p X_{ip}))^2$$

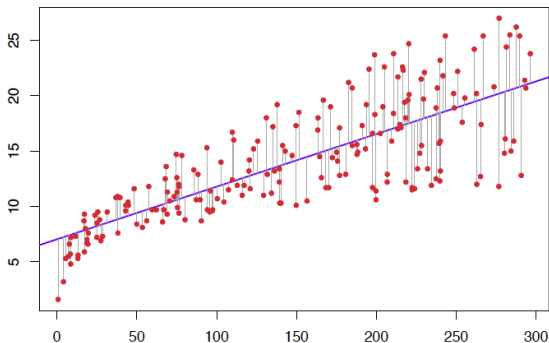
is as small as possible.

- ▶ Equivalently, we're looking for coefficient estimates such that

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

is as small as possible, where \hat{y}_i is the i th predicted value.

Least Squares



- Horizontal axis: predictor
- Vertical axis: response
- Red dots: observations
- Purple line: least squares line

Purple line minimizes sum of squared lengths of the gray lines.

Let's Try Out Least Squares in R!

Chapter 3 R lab
www.statlearning.com