Computational Data Analysis

Introduction

Lars Arvastson and Line Clemmensen

January 31, 2018

Todays Lecture

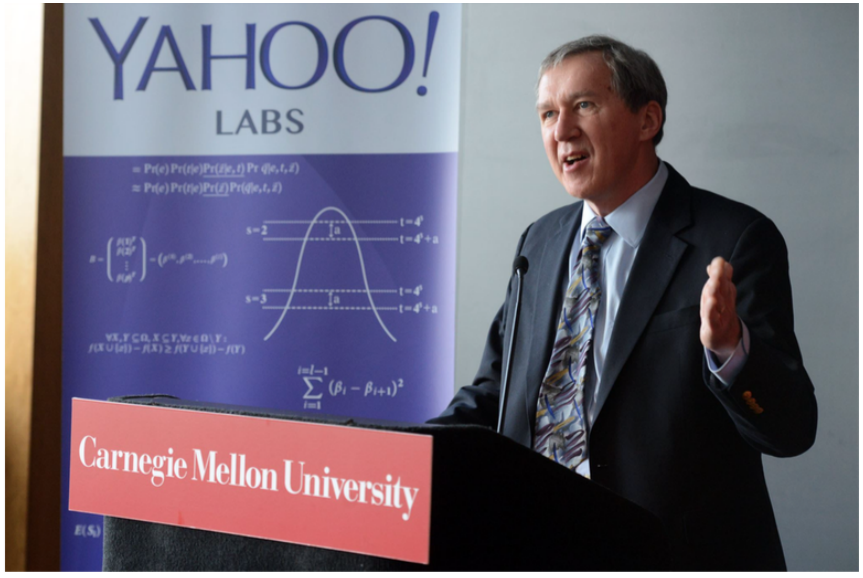
- Computational Data Analysis - What is it?
- Applications - What can it do?
- Course outline - What will we learn?
- Some basic methods and concepts

Machine learning - a definition

Arthur Samuel's definition (1959): "Field of study that gives computers the ability to learn without being explicitly programmed".

Machine learning - a definition

Tom Mitchell's definition (1997): "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*".



Machine learning vs data mining

Machine learning and data mining often employ the same methods and overlap significantly.

Machine learning focuses on prediction, based on *known* properties learned from the training data.

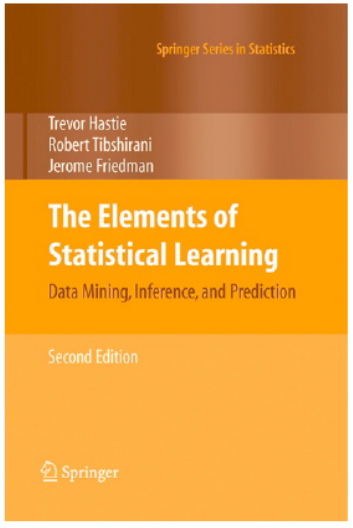
Data mining focuses on the discovery of (previously) unknown properties in the data.

Statistical learning

"This book is about learning from data"

- Trevor Hastie
- Robert Tibshirani
- Jerome Friedman

Stanford University, Department of Statistic



Computational Data Analysis

- **Computational**, to acknowledge the computer science *engineering* aspect.
- **Data analysis** include data mining, data processing, preparation, as well as the statistical analysis/machine learning.
- Data science (a more recent term) *covers all* the mentioned fields.

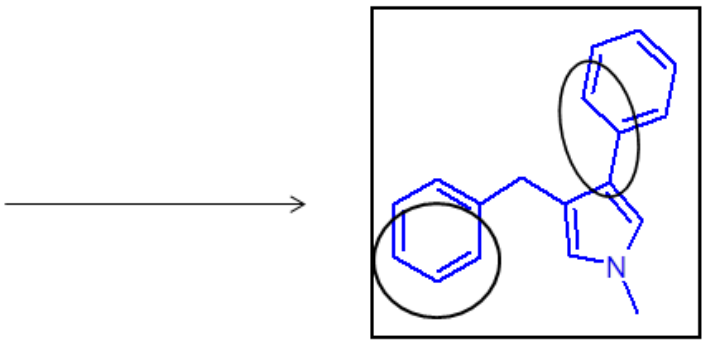
Application - Accident detection

Airbag for cyclists

- Sensors for rotation and acceleration in 3 dimensions.
- Acquire new data 300 times per second.
- Instant decision crash/no-crash.
  - No mistakes allowed.
  - Limited data from accidents.
- Classification with the **Support Vector Machine**.

Application - Life Science - Toxicity

- Database with 5000 molecules
  - Toxicity
  - 200 features describing each molecule
- Experimental descriptors or physico-chemical property descriptors
  - Molecular weight
  - Lipophilicity (logP)
  - Molar refractivity
  - Dipole moment
  - Atomic charges
  - ...
- Symbolic representation descriptors
  - 2D descriptors
  - Count descriptors (atoms, bonds,...)
  - Fingerprints (structural fragments)
  - ...

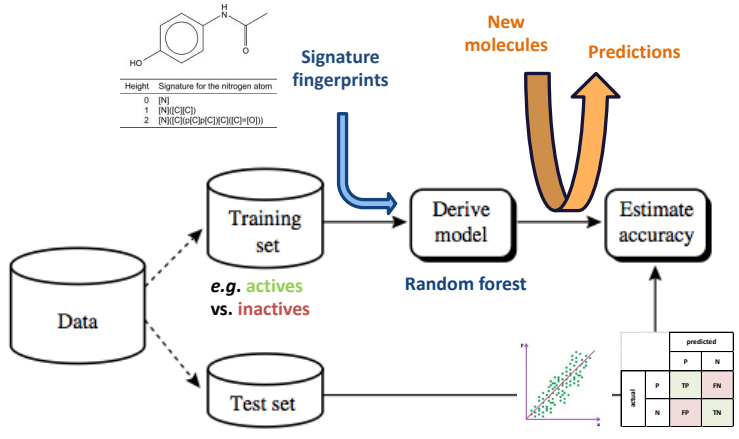


Application - Life Science - Toxicity

Associating molecule features with toxicity using **Random Forest**

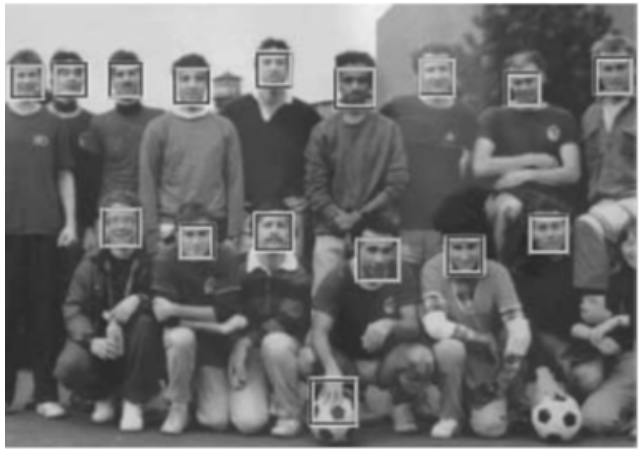
Knowledge based predictive models

Domain	*Property*	*Type*
Absorption	Permeability	Classifier
Distribution	Pgp Efflux	Classifier
Metabolism	Microsomal stability	Classifier
	hERG	Classifier
Toxicity	Cytotoxicity	Classifier
PhysChem	Solubility	Numerical



Application - Face detection

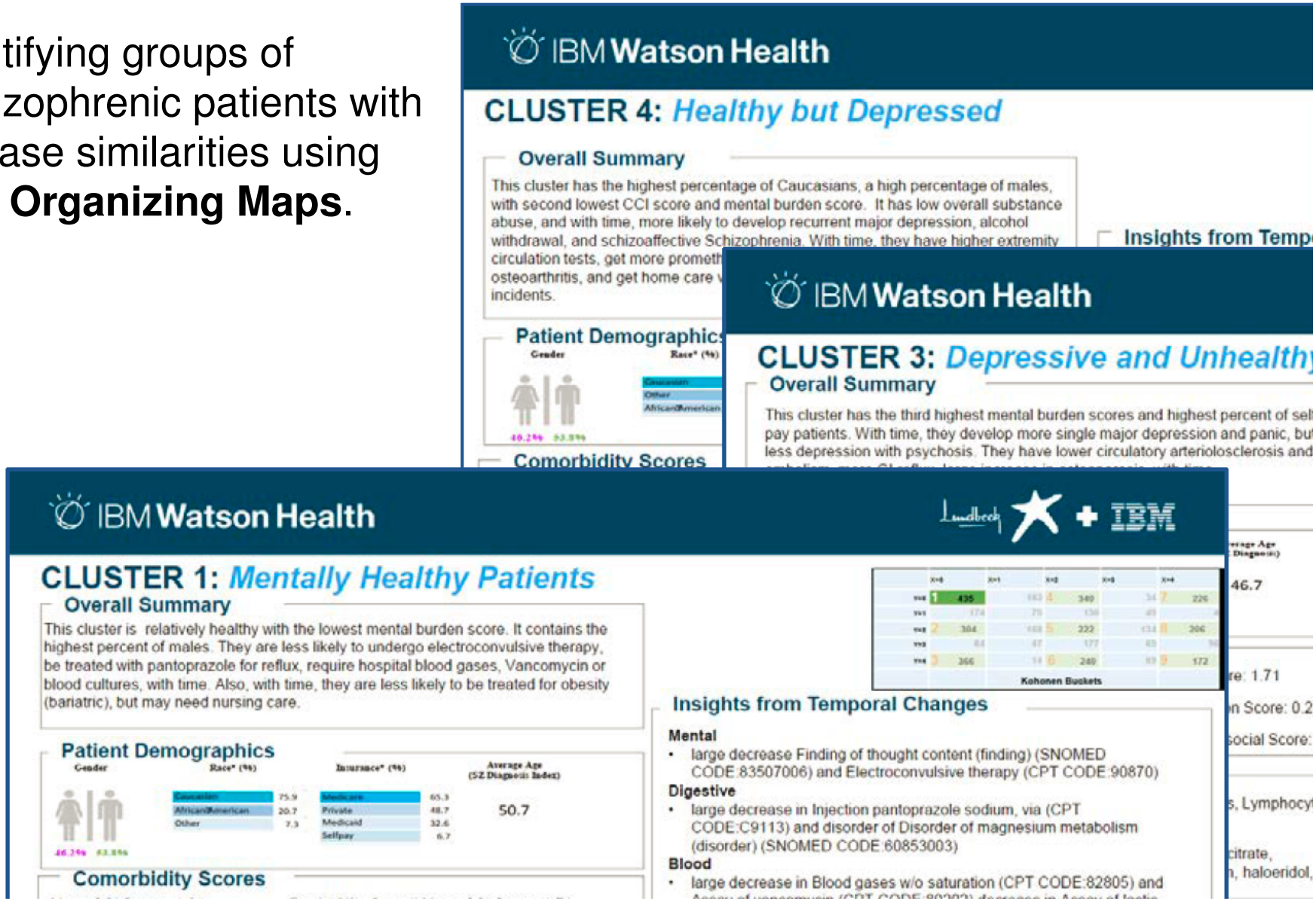
- Face detection using **AdaBoost**
- Algorithm trained on a data set with pictures of faces and random pictures.
- Used in pocket cameras, cell phones, facebook etc.
- Notice the football!



P. Viola and M. Jones; International Journal of Computer Vision, 2004

Application - Patient Segmentation

Identifying groups of Schizophrenic patients with disease similarities using **Self Organizing Maps**.



Application and opportunities everywhere

Banking/Finance	Risk and credit analysis Customer segmentation High-frequency trading
Government	Anti-terrorism efforts Profiling tax-cheaters
Insurance	Setting rates for premiums Fraud detection
Internet	Smart search engines, web marketing Spam filters
Marketing	Segmentation and customer profiling
Life-science	Diagnostics Predicting drug response Identify targets and promising molecules

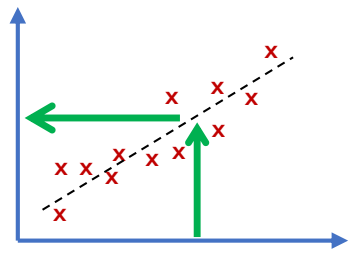
Course goals

- Get experience with a selection of modern statistical tools for accurate data analysis.
  - Emphasis on life-science and industrial applications.
- Statistical engineering
  - Solve unstructured problems

Types of problems

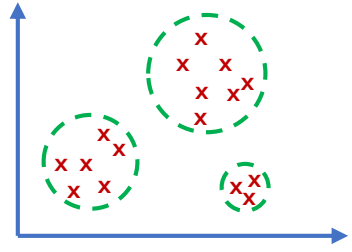
Supervised learning:

- The computer is presented with example inputs and their desired outputs.
- Relies on valid label assignments and a useful response.



Unsupervised learning:

- No labels are given to the learning algorithm.
- The computer learns a structure from the input data.
- Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.
  - Feature generation
  - Outlier detection



The learning cycle

- Find a hypothesis (data driven)
- Test a hypothesis (hypothesis drive)

Course overview

Lecture	Date	Subjects	Lecturer	Literature
1	1/2	Introduction to Computational Data Analysis [OLS, Ridge, Fischer LDA, KNN]	Lars, Line	ESL Chapters 1, 2, 3.1, 3.2, 3.4.1, 4.1 and 13.3
2	8/2	Model Selection [CV, Bootstrap, Cp, AIC, BIC, ROC]	Line	ESL Chapter 7 and 9.2.5. You may safely skip sections 7.8 and 7.9
3	15/2	Sparse Regression [Lasso, Elastic Net] Case 1 presentation	Line	ESL Chapter 3.3, 3.4 and 18
4	22/2	Linear Classifiers and Basis Expansion [LDA, QDA, Logistic regression, Splines]	Lars	ESL Chapter 4.3, 4.4, 5.1 and 5.2
5	1/3	Support Vector Machines and Convex Optimization	Lars	ESL Chapter 4.5, 12.1, 12.2 and 12.3.1
6	8/3	Sub-Space Methods [PCA, CCA, PCR, PLS]	Line	ESL Chapter 14.5.1, 14.5.5 and 3.5
7	15/3	Unsupervised Clustering [Hierarchical clustering, K-means, GMM, Gap-statistics]	Lars	ESL Chapter 14.3

Course overview, cont'd

Lecture	Date	Subjects	Lecturer	Literature
	20/3	**Deadline for handing in Case 1** - short report		
8	22/3	Classification And Regression Trees Discussion of Case 1 and case competition	Lars	ESL Chapter 9.2
	29/3	Easter Holiday		
9	5/4	Ensemble Methods [Bagging, Boosting and Random Forest]	Line	ESL Chapter 8.7, 10.1 and 15
10	12/4	Unsupervised Decomposition [SC, NMF, AA, ICA]	Morten	ESL Chapter 14.6, 14.7. Article "Sparse Coding" Nature
11	19/4	Multi-Way Models	Morten	WireOverview.pdf
12	26/4	Artificial Neural Networks and Self Organizing Maps	Lars	ESL Chapter 11.1-11.5 and 14.4
	1/5	**Deadline for handing in Case 2** - poster pdf		
13	3/5	Case 2 poster presentation, Industrial examples, Principal Data Scientist Sune Askjær presenting Think Big Analytics		
	18/5	**Oral examination** (plus more days that week)		

Examination and mandatory projects

- **Oral examination**
- **Must pass** the two cases to go to the exam.
- Notice the **deadlines** for the cases.
- Grade based on the **exam only**.

Course material

Slides, notes and exercises

- Defines the course
- Our selection of topics
  - Broad overview
  - Theoretical level
  - Understanding

The Elements of Statistical Learning

- Book by T. Hastie, R. Tibshirani and J. Friedman (2^{nd} edition)
- Available online (10^{th} edition)
- Fairly statistical angle
- Supplementary/suggested papers and book chapters will be given

Computer exercises and cases

- Theory and application - hands on
- Software: **Matlab** (and we try to give **R** and **Python** as well).

Our aim

After the course you should...

- Say that it was **tough** but you learned a lot.
- Have a **toolbox** of methods to use.
- Know the methods such that you can,
  - **demonstrate** understanding of their usage,
  - **explain** methods to others,
  - **motivate** choices of methods and parameters.
- Be **ready** to learn more.

About us

Line Clemmensen

Associate Professor
Statistics and Data Analysis, DTU

- Industrial optimization and AI
- Machine learning in Life-science
- Learning from small sample sizes
- Predictive analytics in Industry

Principal Data Scientist, Mærsk Digital
PhD in Statistical Image Analysis

lkhc@dtu.dk

About us

Lars Arvastson

External lecturer

Senior Specialist
Bioinformatics, H. Lundbeck A/S

- Biomarkers
- EEG signal processing
- Drug target screening models
- Drug development

PhD in Mathematical Statistics
Lund University, Sweden

larv@lundbeck.com

About us

Morten Mørup

Associate Professor
Cognitive Systems, DTU

PhD thesis on Unsupervised Learning

Research interests:

- Brain imaging: Modeling EEG and fMRI signals
- Multi-media: Modeling audio and media data
- Complex networks: Structure identification

mmor@dtu.dk

About us

David Norsk

PhD student
Image Analysis and Computer Graphics, DTU Compute

dnor@dtu.dk

Sunna Lilja Björnsdóttir

MSc student
Applied Mathematics and Computer Science, DTU Compute

s161622@student.dtu.dk

Helga Svala Sigurðardóttir

MSc student
Applied Mathematics and Computer Science, DTU Compute

s163507@student.dtu.dk

About you

- What is your background?
- Why did you choose this course?
- What do you expect from us?
- What should we expect from you?

Today's lecture

- Terminology
- Basic regression
- Basic classification

Terminology - a suggestion

The **data matrix** X. (Capital letters denote matrices)

- Rows correspond to **samples**
- Columns correspond to **variables**
- Size of X is $(n \times p)$, n samples and p variables.

The **response variable** y (small letters denote vectors)

- Usually a $(n \times 1)$ vector
- Also known as the **output**

The **model coefficients** β

- A $(p \times 1)$ vector

The **model error** vector e

- An $(n \times 1)$ vector

The **prediction error** vector ϵ

- An $(n \times 1)$ vector
- Also known as the **residual**

A **regression model** that is linear in β can be written
$y = X\beta + e$

Inputs, predictors, features...

What we call our variables differs with who we are,

Machine learning	**Statistics**	**Pattern recognition**
X inputs	X predictors	X features
Y outputs	Y responses	Y responses

Terminology

Methods are divided into categories depending on the nature of the response variable.

No response variable available - **unsupervised**

- Principal Component Analysis
- Cluster analysis

Response variable(s) available - **supervised**

- Categorical response - **classification**
  - Linear Discriminant Analysis
  - K Nearest Neighbors
  - Support Vector Machine
- Continuous response - **regression**
  - Ordinary Least Squares
  - Ridge regression
  - K-Nearest-Neighbors

Basic regression

- Ordinary Least Squares
- Bias-variance
- Expected Prediction Error
- Ridge regression

Review - Linear regression

Given a continuous response measurement, find the relation of this variable to a set of input variables.

Model: $y = X\beta + e$

Example

- Predict ice-cream sale, y, based on measurement of temperature, x_1, and sunshine hours, x_2, $X = [x_1, x_2]$. The relation is given by β.
- Predict house prices based on interest rate, unemployment rate, region, size and building year.
- Predict life expectancy given smoking habits, age, BMI etc.

Ordinary Least Squares, OLS

Find the β that minimizes the residual error, $y - X\beta$. Positive and negative errors are equally bad and large errors are worse than small errors. Hence, minimize

$$||y - X\beta||_2^2 = \sum_{i=1}^{n}(y_i - X_i\beta)^2$$

First done by Carl Friedrich Gauss in 1794

- First used in 1801
- Published 1809

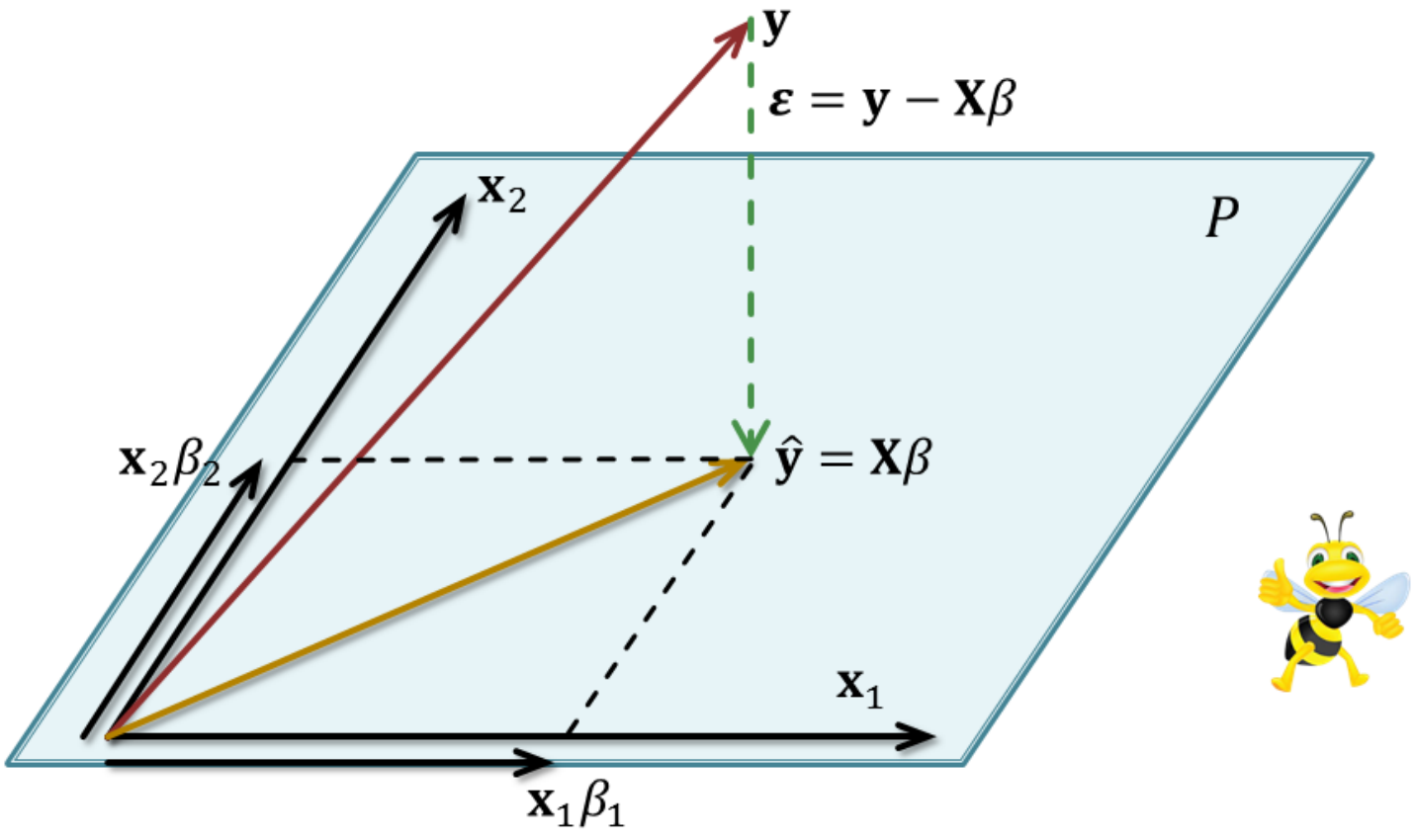
$$\beta_{OLS} = \arg\min_{\beta} ||y - X\beta||_2^2$$

$$\frac{\partial}{\partial\beta}(y - X\beta)^T(y - X\beta) = 0$$

$$-2X^T(y - X\beta) = 0$$

$$\beta_{OLS} = (X^TX)^{-1}X^Ty$$

OLS as orthogonal projection



Choose β such that the residual is orthogonal to the hyperplane, ie shortest distance from point to plane.

Use the fact that x_1 and x_2 are both orthogonal to $y - X\beta$ to derive the OLS solution.

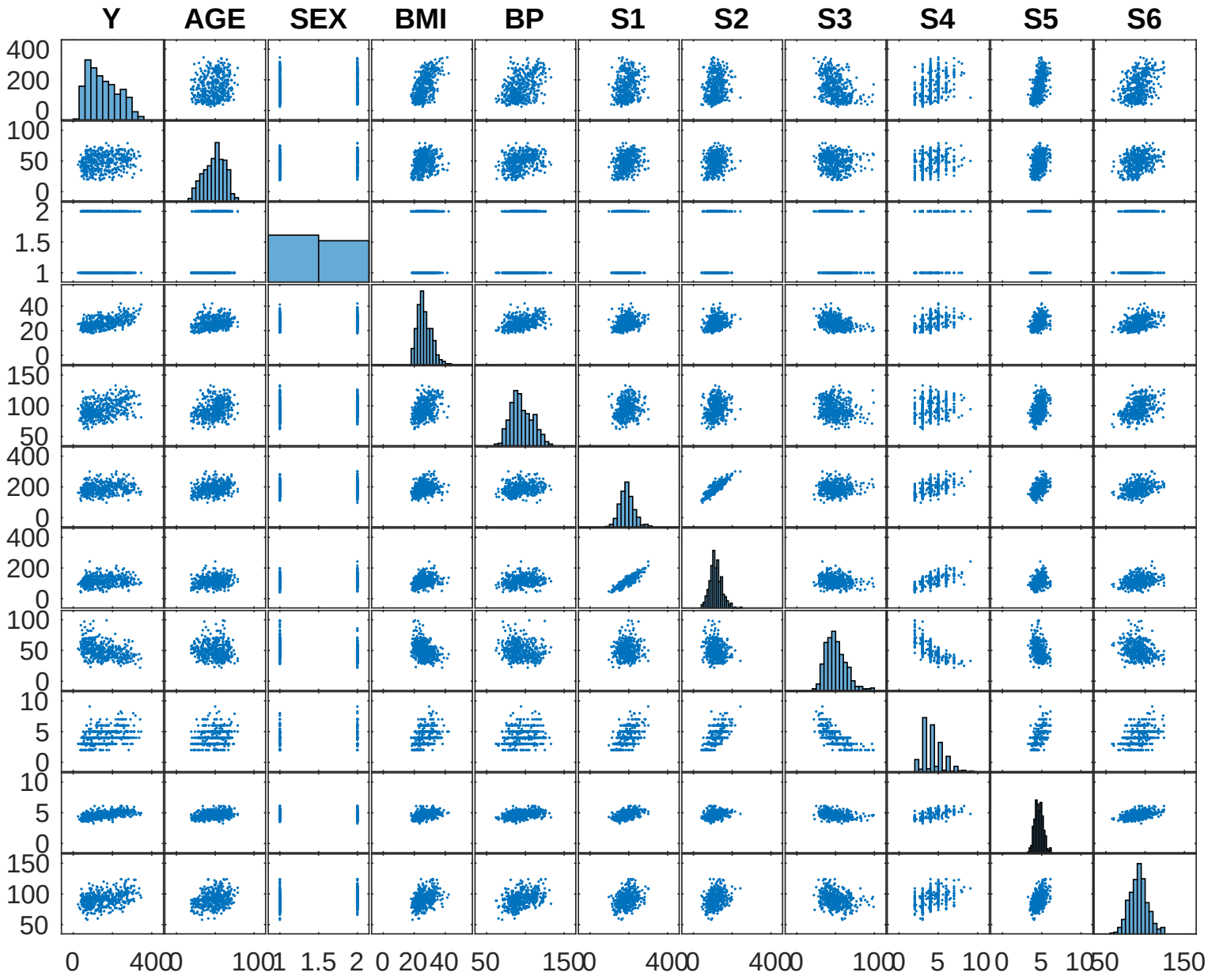
Example: Diabetes

- 442 observations on 10 variables.
- "Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline."
- Data from Efron et al. "Least Angle Regression" 2004

Next page shows a scatterplot of data. Generated in Matlab by

```
T = readtable('DiabetesData.txt');
Y = T{:,11};
X = T{:,1:10};
plotmatrix([Y,X]);
```

Diabetes scatterplot



Diabetes linear regression model

```
>> lme = fitlme(T,'Y ~ AGE+SEX+BMI+BP+S1+S2+S3+S4+S5+S6')
```

Name	Estimate	SE	pValue
'(Intercept)'	-334.57	66.61	7.4725e-07
'AGE'	-0.036361	0.21432	0.86536
'SEX'	-22.86	5.7627	8.5294e-05
'BMI'	5.603	0.70813	2.1508e-14
'BP'	1.1168	0.22242	7.5301e-07
'S1'	-1.09	0.56615	0.054853
'S2'	0.74645	0.52419	0.15517
'S3'	0.372	0.77267	0.63044
'S4'	6.5338	5.884	0.26743
'S5'	68.483	15.474	1.2187e-05
'S6'	0.28012	0.26989	0.2999

Two reasons for linear regression modelling:

Insight: Tell which variables that are important.

Prediction: Predict outcome, $\hat{y}$, from an input variable x, $\hat{y} = x\hat{\beta}$.

Exercise, diabetes marginal effects

Previous p-values are calculated assuming dropping one variable while keeping all others in. Two very correlated variables could therefor show weak p-values even if they are important.

Your task is to calculate p-values looking at one variable at a time (one linear model for each variable)

Name	Estimate	SE	pValue
'(Intercept)'	_____	_____	_____
'AGE'	_____	_____	_____
'SEX'	_____	_____	_____
'BMI'	_____	_____	_____
'BP'	_____	_____	_____
'S1'	_____	_____	_____
'S2'	_____	_____	_____
'S3'	_____	_____	_____
'S4'	_____	_____	_____
'S5'	_____	_____	_____
'S6'	_____	_____	_____

Exercise, diabetes reduced model

We have just seen that all variables have strong link to the output when analyzed on its own.

Now, build a reduced model by starting with the full model eliminating the least significant variable until all variables in the model are significant.

Name	Estimate	SE	pValue
'(Intercept)'			

Interpreting regression coefficients

Correlated predictors are a **problem**

- The variance of the estimates tends to increase.
- Interpretation becomes hazardous. When x_j changes everything else changes as well.

Ideally we have predictors that are uncorrelated.

- Often we need a **designed experiment** to obtain this.
- That is when coefficients can be estimated and tested separately.
- Interpretation is easy. A unit change in x_j causes a change of β_j in Y, holding all other variables fixed.

Regression for prediction

Regression can be done to obtain a prediction model $\hat{y}(x) = x\hat{\beta}$

- The variance of the prediction is $Var(\hat{y}(x)) = xCov(\hat{\beta})x^T$. The order of the variance is $\mathcal{O}(\frac{p}{N}\sigma^2)$. **More variables** means more uncertainty and **more observations** means less uncertainty.
- **Removing variables** can decrease variance but it comes at a price. There will be a **systematic error** (bias) due to the missing variables.

Properties of the OLS

We can characterize a model in terms of its **bias** and **variance**.

Methods in this course often aim to **lower variance** - at the price of **increasing bias**.

Bias

What is it?

- The difference between an expected value and the true value.

Bias of what?

- Could be of β the model parameters.
- Could also be of the predictions $\hat{y}$.

Unbiased?

- Repeat experiment, take average of β or $\hat{y}$.
- Will be equal to the true value or unbiased.

Examples on the blackboard...

Variance

We might be right on average (unbiased) but we only do one experiment.

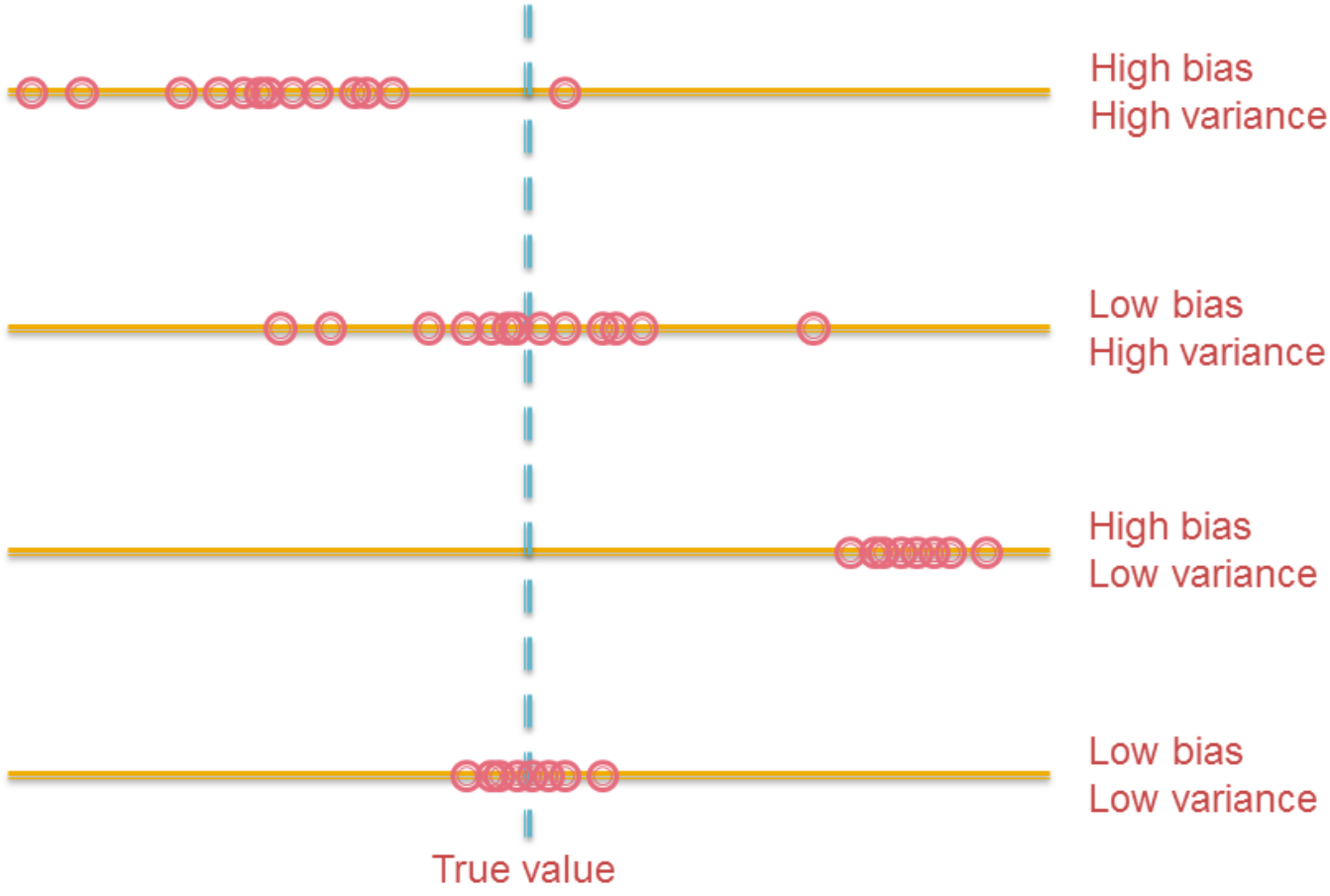
How far are we from the true value?

- Depends on the variance.

- **High variance,** we might end up far from the true value.
- **Low variance,** we get almost the same result every time, how far it is from the true value depends on the bias.

Examples on the blackboard...

Illustrations



Properties of OLS

Ordinary least squares (OLS) is great!

OLS is the **best linear unbiased estimate** (BLUE)

- Unbiased: $E(\beta_{OLS}) = \beta$
- Best unbiased: $Var(\beta_{OLS}) \leq Var(\beta_{linear})$

Why use anything else?

1. What happens, with $\beta_{OLS} = (X^T X)^{-1} X^T y$, if $p > n$?

2. Performance measurement for regression methods:
   Expected Prediction Error (EPE)

$$EPE(x_0) = \lim_{k \to \infty} \frac{1}{k} \sum_{i=0}^{k} ||y_i(x_0) - \hat{f}(x_0; D_i)||^2$$

- The expected squared prediction error if we repeat the experiment, including data gathering, many times
- Training data D_i are selected at random.
- A smaller EPE value is better!

Expected Prediction Error

Three sources of errors contribute to the EPE,

$$\begin{aligned} EPE(x_0) &= E_{y,D|x_0}||y(x_0) - \hat{f}(x_0; D)||^2 \\ &= \sigma_e^2 + \text{bias}^2(\hat{f}(x_0; D)) + \text{variance}(\hat{f}(x_0; D)) \end{aligned}$$

1. irreducible error $\sigma_e^2 = E_y\left(y(x_0) - f(x_0)\right)^2$
2. $\text{bias}^2(\hat{f}(x_0; D)) = \left(E_D(\hat{f}(x_0; D)) - f(x_0)\right)^2$
3. $\text{variance}(\hat{f}(x_0; D)) = E_D\left(\hat{f}(x_0; D) - E_D(\hat{f}(x_0; D))\right)^2$

where $\hat{f}(x_0)$ is the prediction of $f(x_0)$ with observed $y = f(x_0) + e$, $E(e) = 0$, $\text{variance}(e) = \sigma_e^2$ and D is training data.

Exercise

Show that

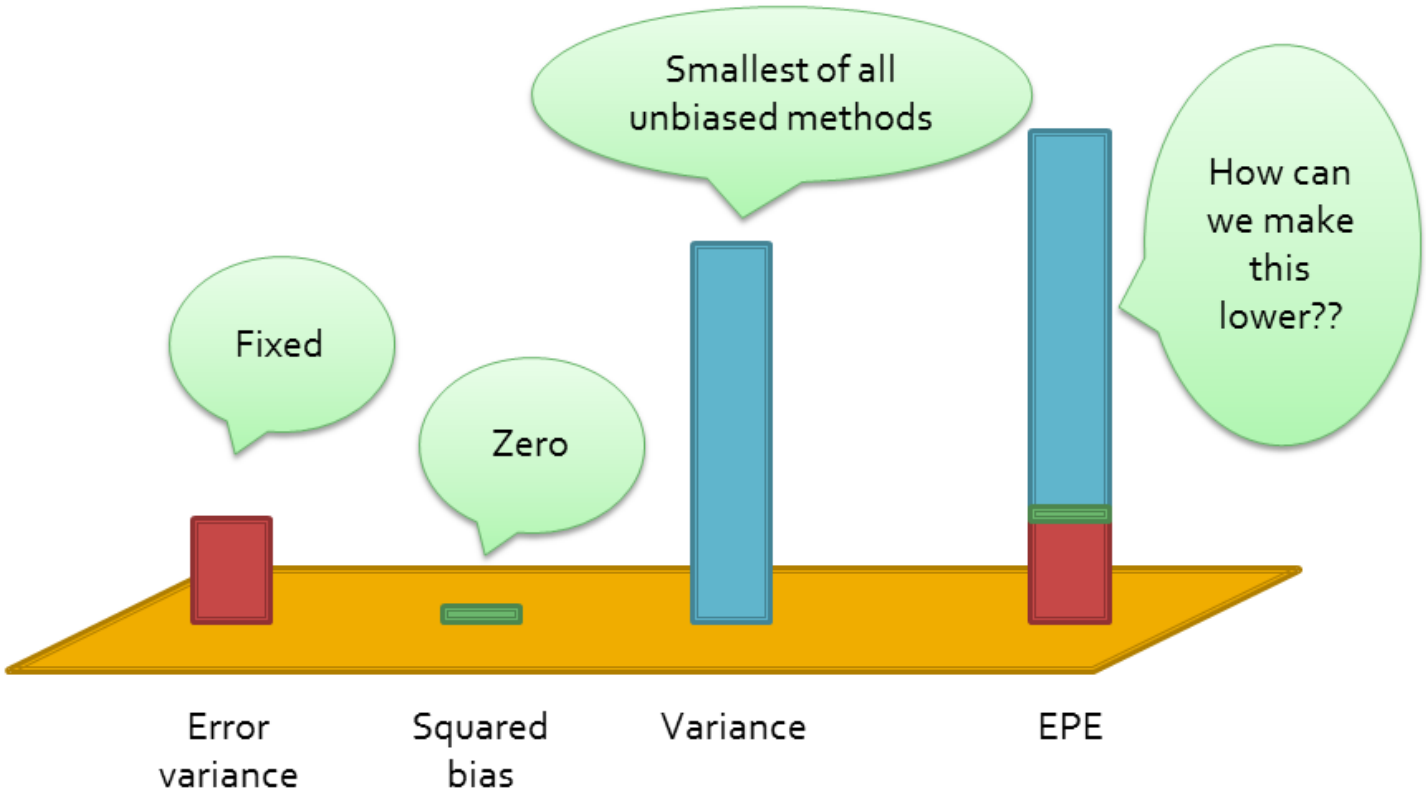
$$E(y - \hat{f})^2 = \sigma^2 + (E(\hat{f} - f))^2 + (E(\hat{f} - E\hat{f}))^2$$

given

$$y = f + e, \quad E(e) = 0 \text{ and } E(e^2) = \sigma^2$$

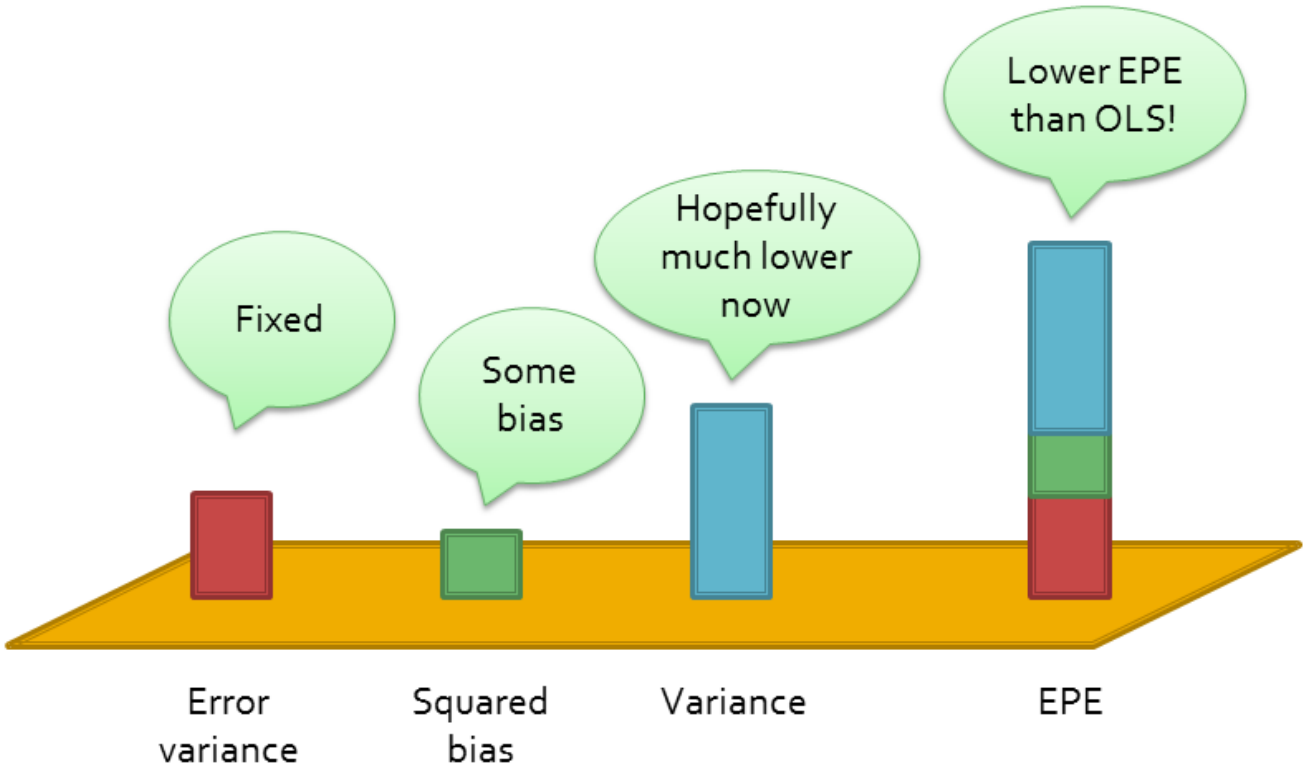
Expected Prediction Error

With OLS,



Expected Prediction Error

Introducing bias,



Making things worse ... and better

- But wait, OLS minimizes the residual
  - How can any method be better?
- Estimated prediction error is measured on **test data**. The residuals from before was measured on **training data**.
- Some models can be expected to perform better on new data - from the real world.
  - Even though they perform worse on the data we are given.

Ridge Regression

1. We wish to lower the variance of $\hat{y} = X\beta$.
2. Lowering the size of β will lower the variance of $\hat{y}$.
3. Lower the size of β by **shrinkage**,
   - $\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} ||y - X\beta||^2 + \lambda||\beta||^2$
   - OLS criterion plus extra term.
   - λ controls the amount of shrinkage.

Ridge regression

Ridge regression has a closed-form solution!

$$\beta_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

- You will derive this expression for β_{ridge} in the exercises.
- We are adding a small number to the diagonal of the matrix to invert.
  - Hence, the name "ridge" regression.
  - Stabilizes the inverse numerically.
  - Ridge regression solutions are available even when $p > n$!

Ridge Regression - Example

- The diabetes data we used before (this time normalized data)
- 442 observations and 10 variables.
- Response variable describe disease progression one year after baseline.
- Variables include various clinical measure along with age, sex, BMI and blood pressure.

Important: Data are centered (mean of each variable is zero) and normalized (variables scaled such that standard deviation equals 1).

Centering data removes the mean, causes the shrinkage to shrink towards zero.

Normalizing data puts equal importance to all variables due to the penalty term $||\beta||^2$.

Ridge Regression - Example

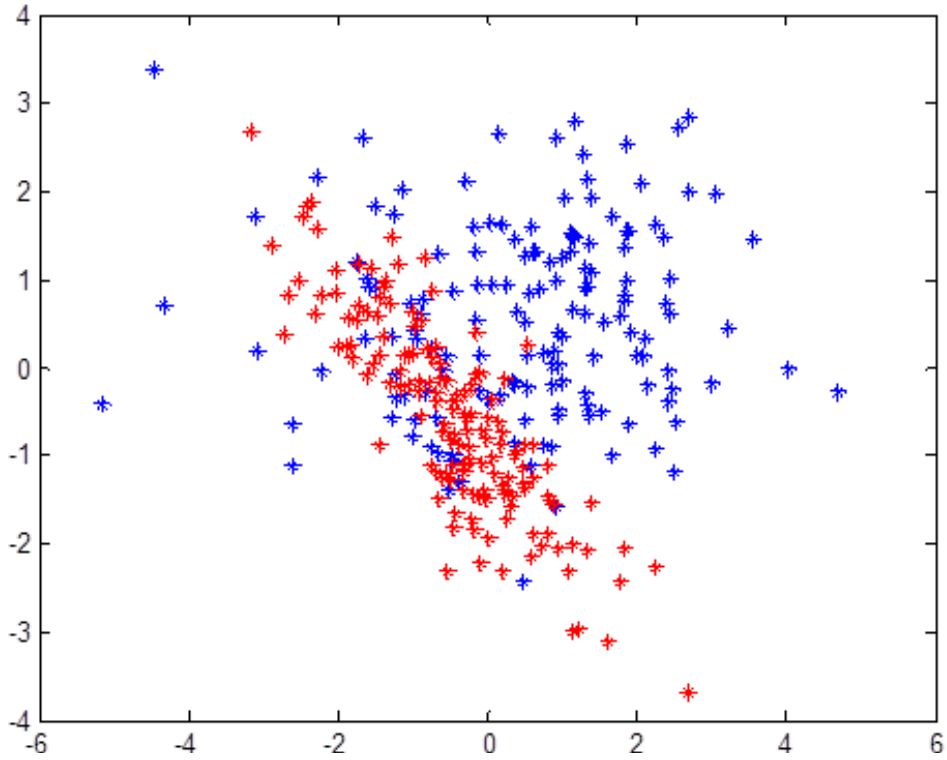
Name	OLS Estimate	Ridge Estimate, $\lambda = 1$	Ridge Estimate, $\lambda = 100$
Age	-0.0062	-0.0056	0.0057
Sex	-0.1481	-0.1472	-0.1094
BMI	0.3211	0.3217	0.2775
BP	0.2004	0.1996	0.1731
S1	-0.4893	-0.3906	-0.0268
S2	0.2945	0.2161	-0.0481
S3	0.0624	0.0189	-0.1165
S4	0.1094	0.0976	0.0743
S5	0.4640	0.4264	0.2421
S6	0.0418	0.0424	0.0615

Basic classification

- Fischer Linear Discriminant Analysis
- K-Nearest-Neighbor classification

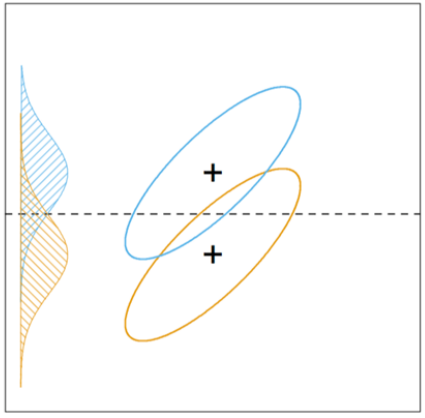
Linear classifiers

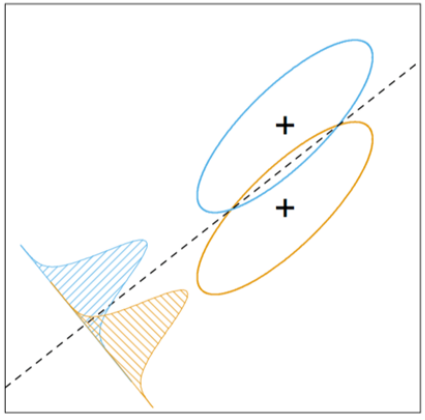
Linear classifiers use a straight line (or a hyperplane in higher dimension) to separate classes.



Fischer Linear Discriminant Analysis

Find a linear combination $Z = a^T X$ such that the between-class variance is maximized relative to the within-class variance.





Maximize between-class variance $a^T \Sigma_B a$

$$\Sigma_B = \sum_{j=1}^{K} (\mu_j - \mu)^T (\mu_j - \mu)$$

Minimize the within-class variance $a^T \Sigma_w a$

$$\Sigma_W = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (X_{ij} - \mu_j)^T (X_{ij} - \mu_j)$$

Group mean is μ_j and total mean is μ. The separating line is given by

$$\max_a \frac{a^T \Sigma_B a}{a^T \Sigma_W a}$$

We will explore this in lecture 4 under a Gaussian assumption.

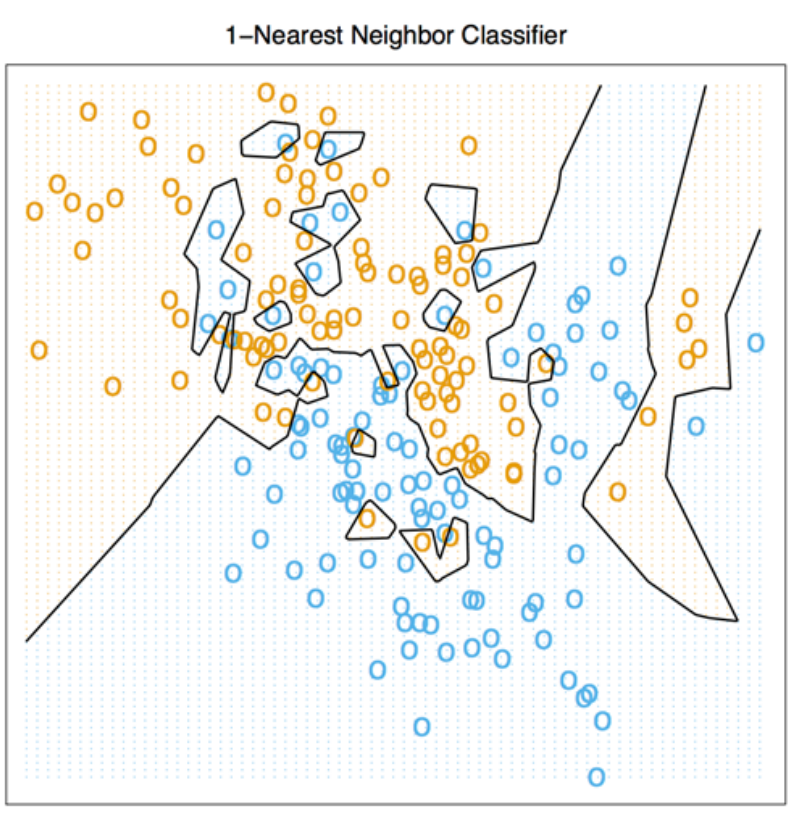
K-Nearest-Neighbor classification

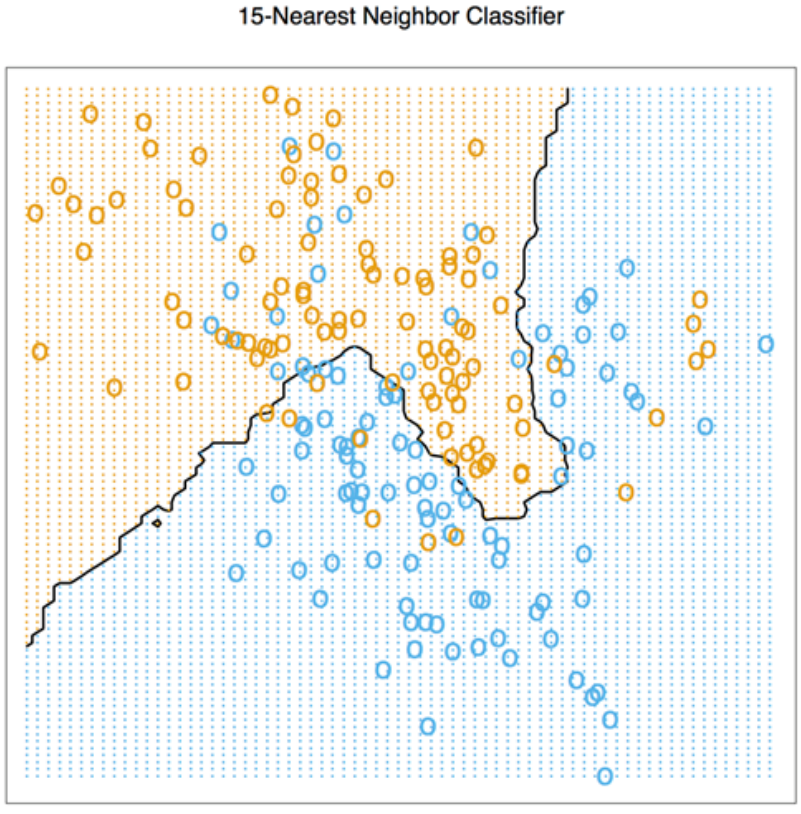
Classify observations according to the majority class of the K nearest neighbors.

- Define distance measure of proximity, eg Euclidean distance.
- It is general practice to standardize each variable to mean zero and variance 1.
- K is a positive integer of your choice. Small values gives low bias, large values will give low variance.

We will learn to find the optimal K for any problem in next lecture!

KNN clustering example





K-Nearest-Neighbors regression

We can use the same technique for regression!

Estimate the response of an observation as the average response of the K nearest neighbors.

- Define distance measure of proximity, eg Euclidean distance.
- It is general practice to standardize each variable to mean zero and variance 1.

Learning objectives

- Regression
  - Ordinary Least Squares (OLS)
    - Unbiased, high variance, $n > p$
  - Ridge regression
    - Biased, lower variance, $p > n$
  - KNN regression
- Classification
  - Fischer LDA
    - Linear
  - KNN classification
    - Flexible
- Bias-variance trade-off and EPE.

Questions?

Today's Exercises

- Solve OLS computationally.
- Examine bias and variance for OLS.
- Derive solution for ridge regression and examine bias and variance.
- Implement and solve KNN regression.