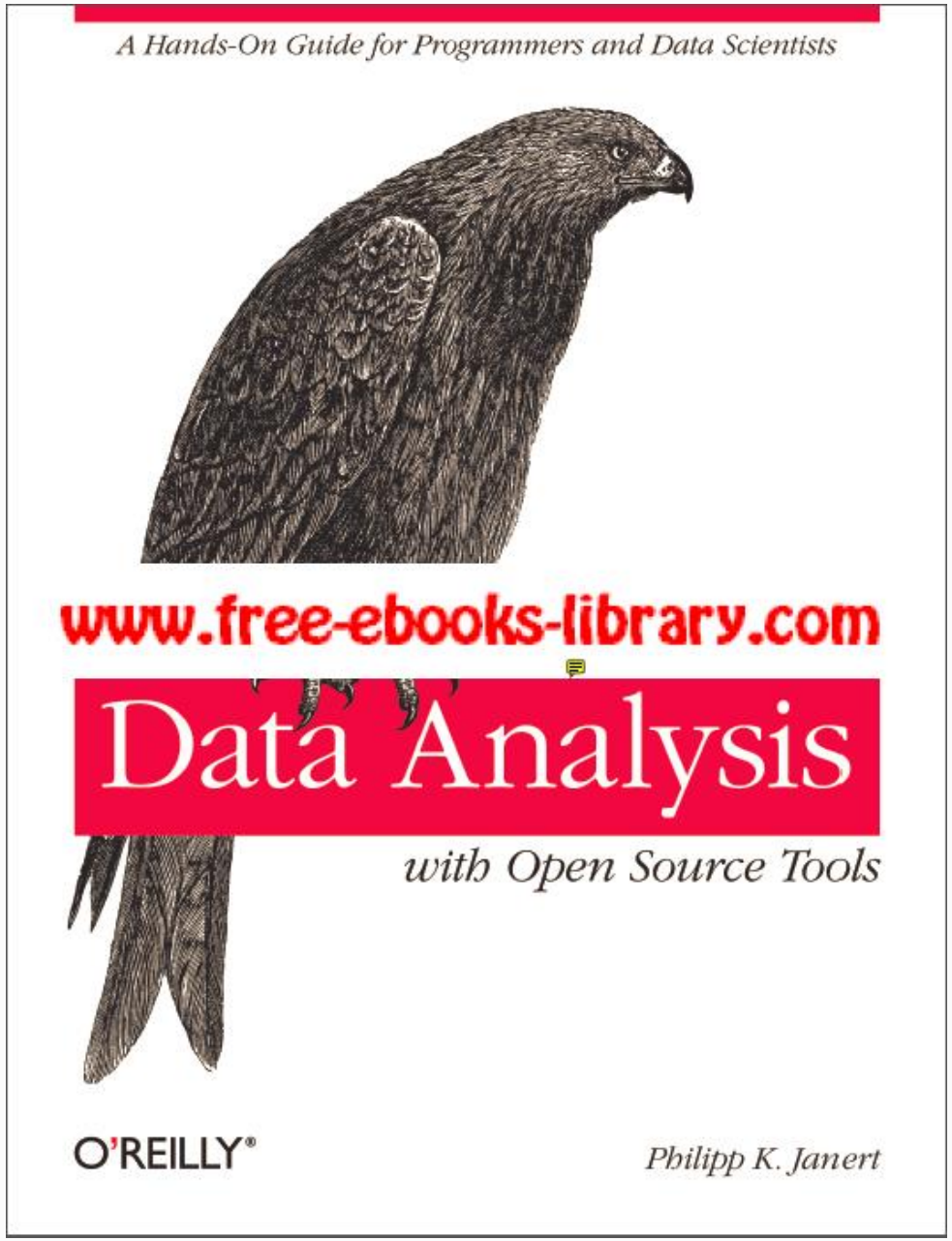
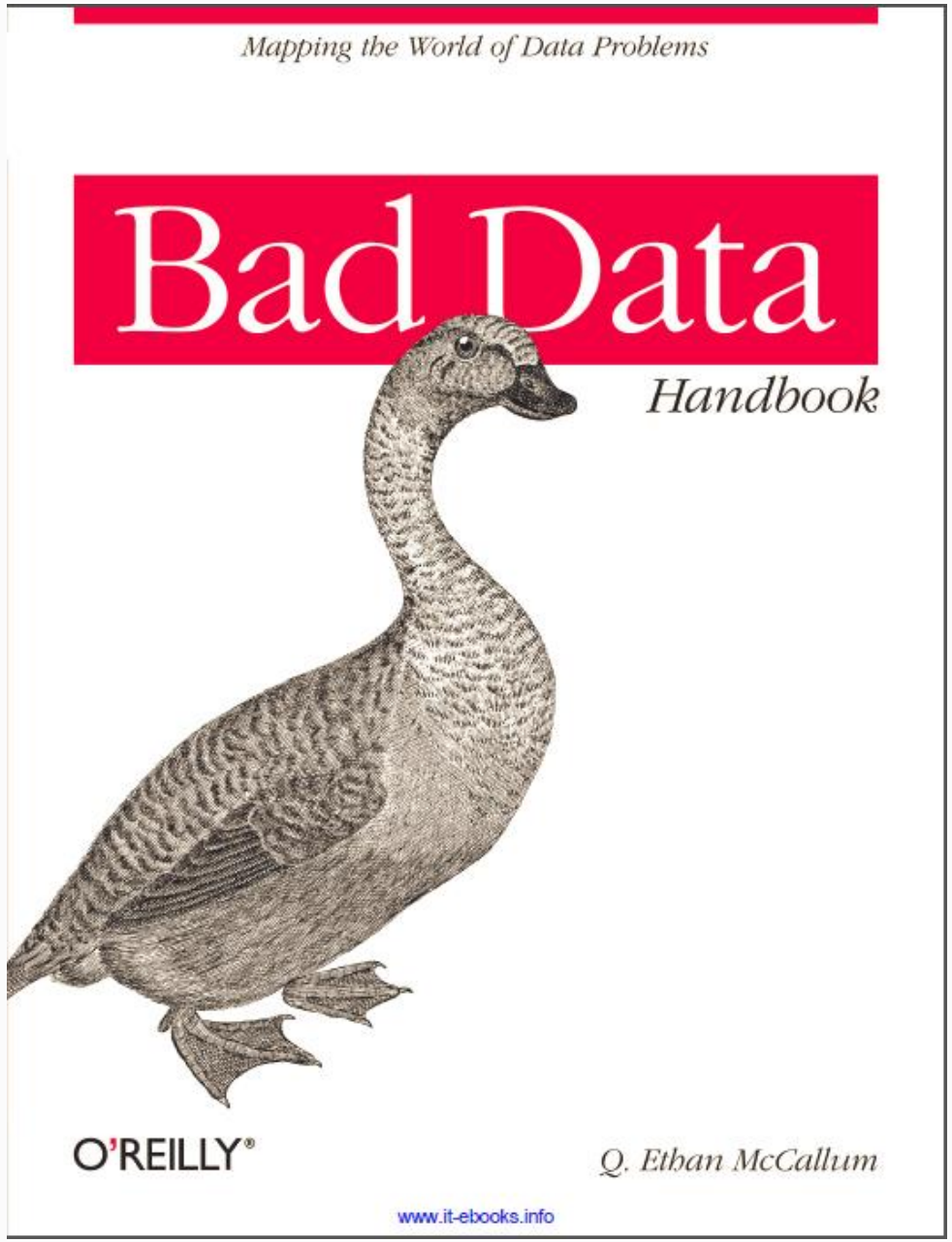
Data Science

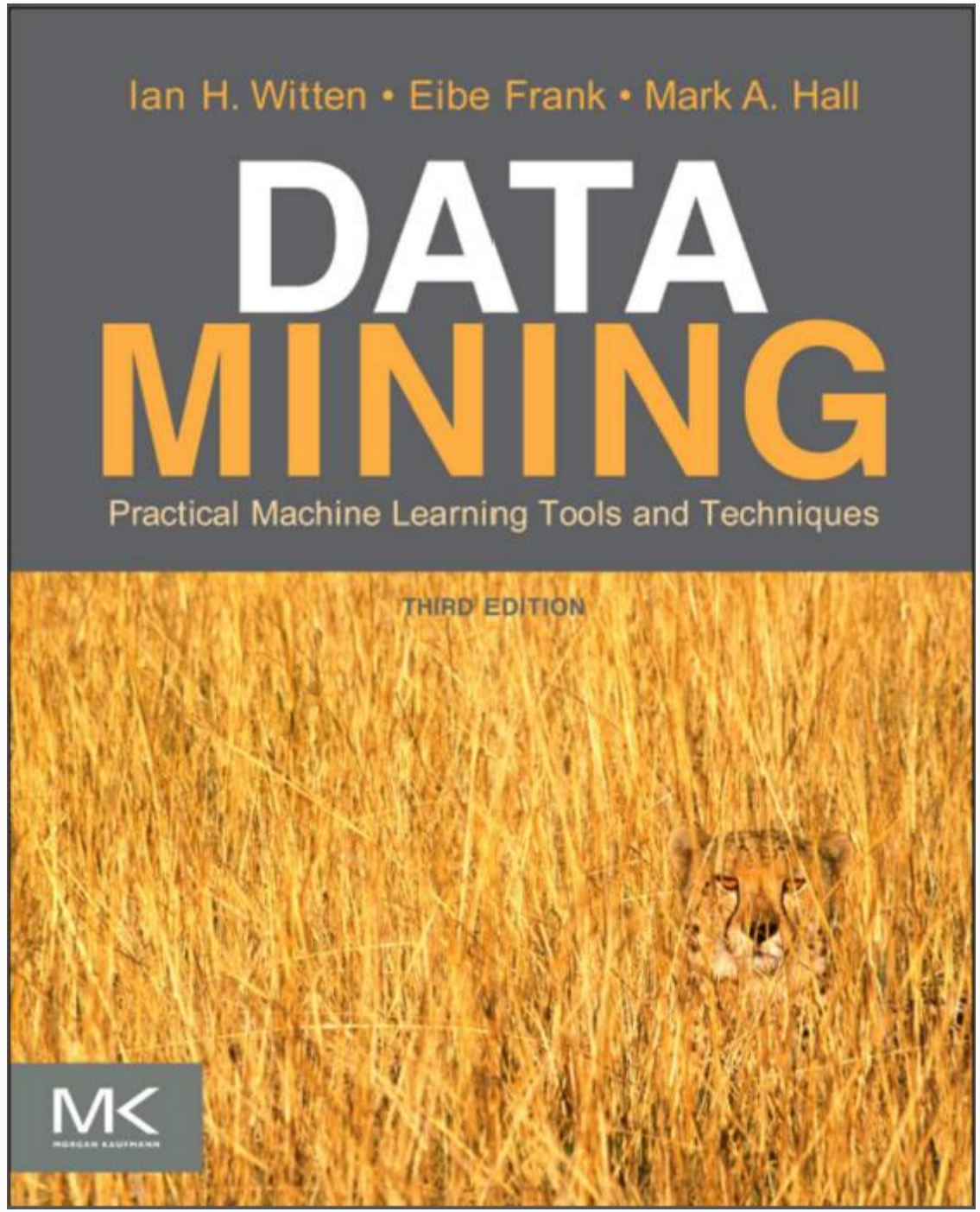
Deriving Knowledge from Data at Scale

Good data preparation is key to producing valid and reliable models…

Nov 16th , 2015


A Hands-On Guide for Programmers and Data Scientists
www.free-ebooks-library.com
Data Analysis
with Open Source Tools
O'REILLY®
Philipp K. Janert


Mapping the World of Data Problems
Bad Data
Handbook
O'REILLY®
Q. Ethan McCallum
www.it-ebooks.info


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK
MORGAN KAUFMANN

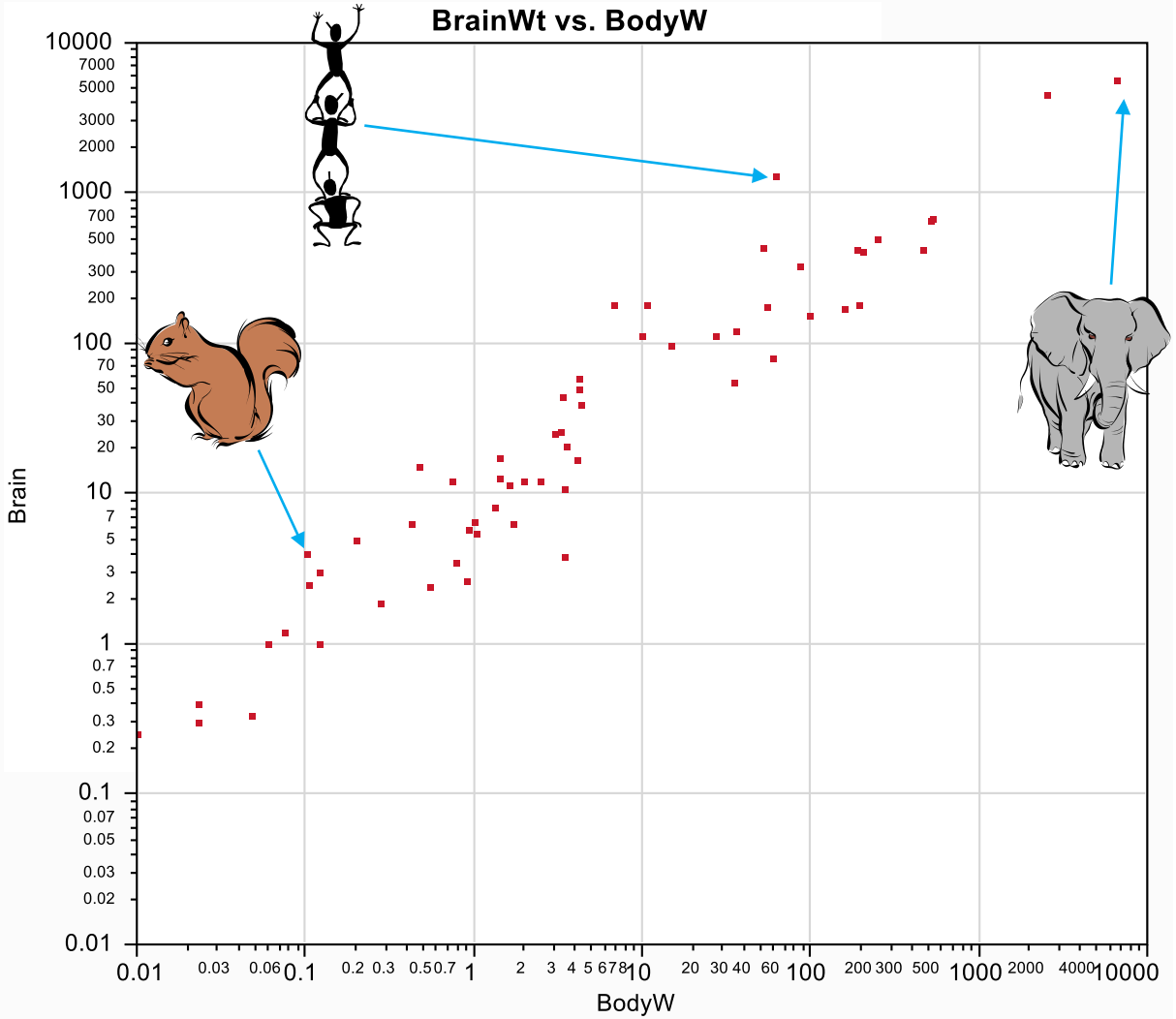
Lecture 7 Agenda

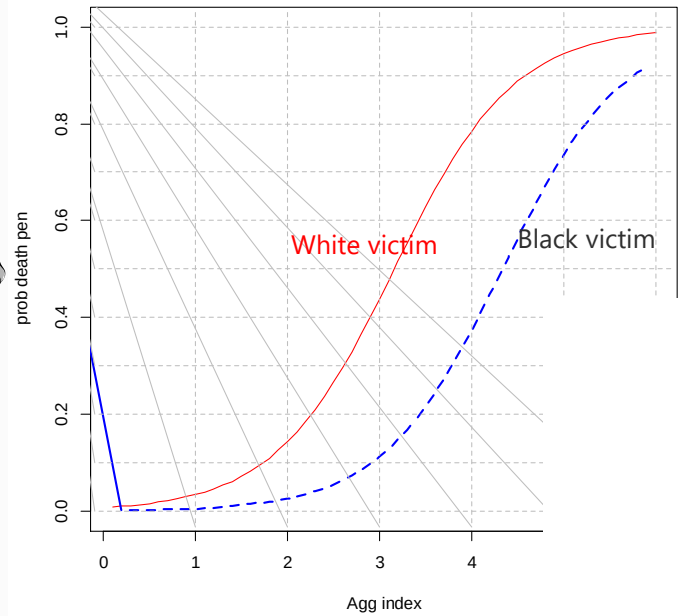
- Regression models
- Support Vector Machines;
- Data Exploration;
- Feature Selection;
- Discuss Class Project;

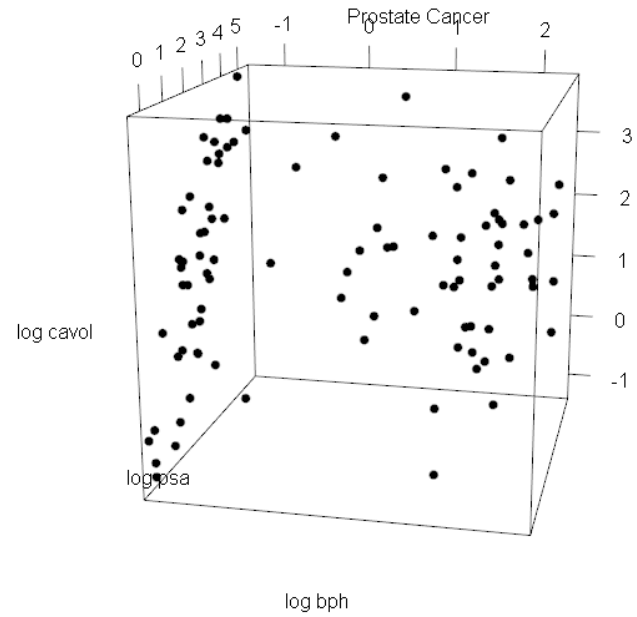
Regression modeling

- Linear regression
- Polynomial regression
- Logistic regression
- Multinomial regression

Some examples







Basic questions

- Is there a relationship? It takes 2 to tango. 2 or more variables?
- What kind of relationship is it? Form? Simple? Complex?
- How strong is the relationship? Clear? Obscure? Amount of scatter?
- Are there anomalies in the relationship? Extreme values? Counter intuitive?
- What is the scope of the relationship? Output range? Input region?
- What could be done from knowing this relationship? Prediction? Assessment?
- Should there be additional information to improve the accuracy of the relationship? Out of scope data available?

Regression model

$$y = f(x) + \varepsilon$$

General regression

Output

f( ) the relationship

Input

Noise,
random error

Output variable:
- dependent variable
- response
- outcome
- variable we want to predict

Input variables:
- independent variables
- contributors
- key drivers
- leading indicators
- variables we use to predict

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Simple linear regression
Very popular!!!

Intercept

slope

Pros/Cons

- Strengths
  - For a local region of X, a linear form is a reasonable approximation.
  - The model estimate is obtained by a "impartial" criterion (least squares).
  - Nice and simple mathematics (easy to illustrate and implement).
  - Foundation for many other methods.

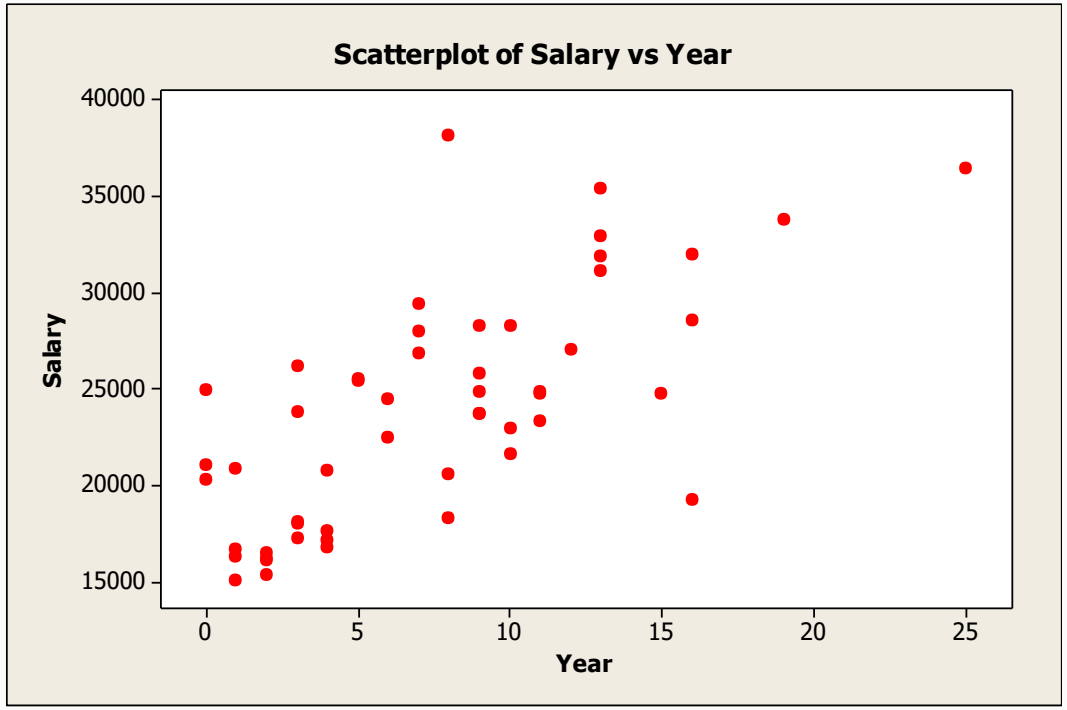
- Limitations
  - Straight line model is not reliable for extrapolation beyond the data scope.
  - Sensitive to extreme values, in Y, in X or in both.

Example 1

A human resource department was interested in attributes affecting employee's salary. The following chart displays a potential relationship between years of service within a company and salary.

Is there a relationship between salary and years of service?

What kind of relationship is it?



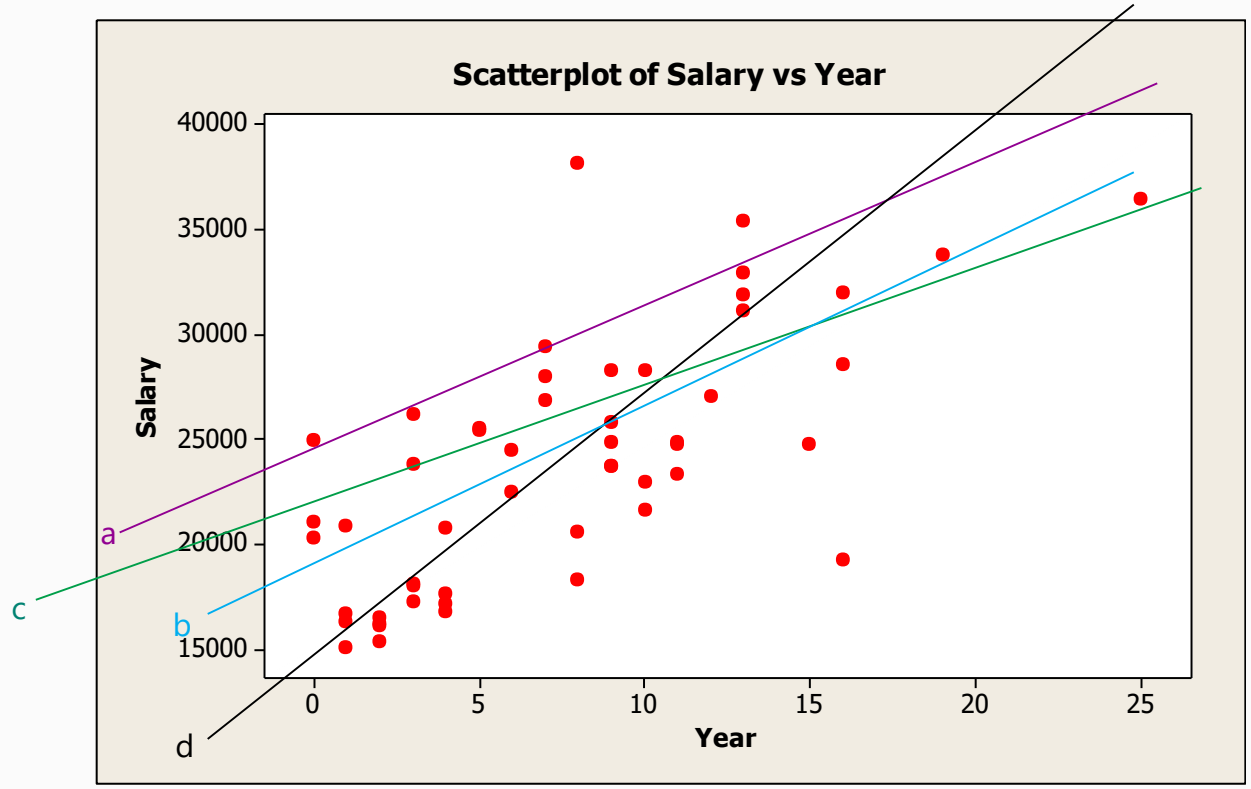
Example 1 continue

A human resource department would like to use that relationship to estimate the salary range for hiring a senior person with 7 years of experience from a competing company.

- Is there a relationship between salary and years of service?
- What kind of relationship is it?

How should one estimate the relationship?

Line a
Line b
Line c
Line d



Example 1 continue

How should one estimate the relationship?

Least square criterion:

A "good" estimate of model (line) is one that minimizes overall squared deviations between the data points and the line.

squared deviations

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

line

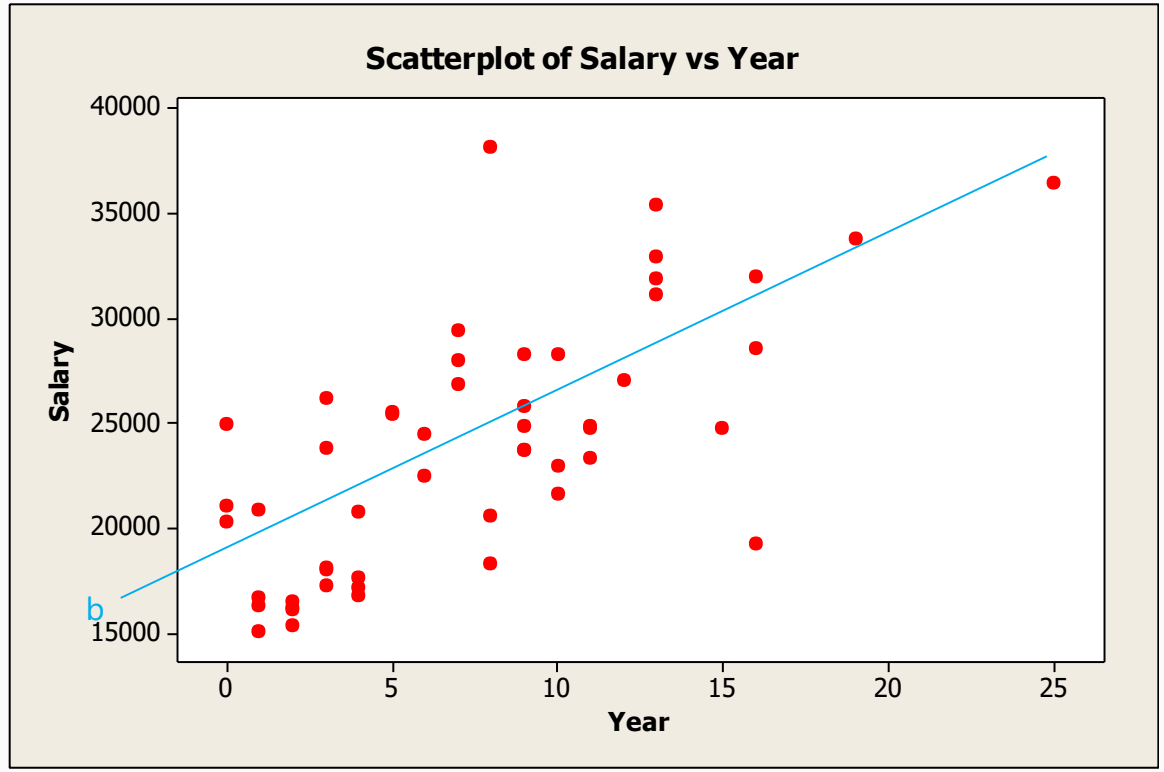
model estimate

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

parameter estimates

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$



Terminology

response variable
(dependent variable)

explanatory variable
(independent variable)

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

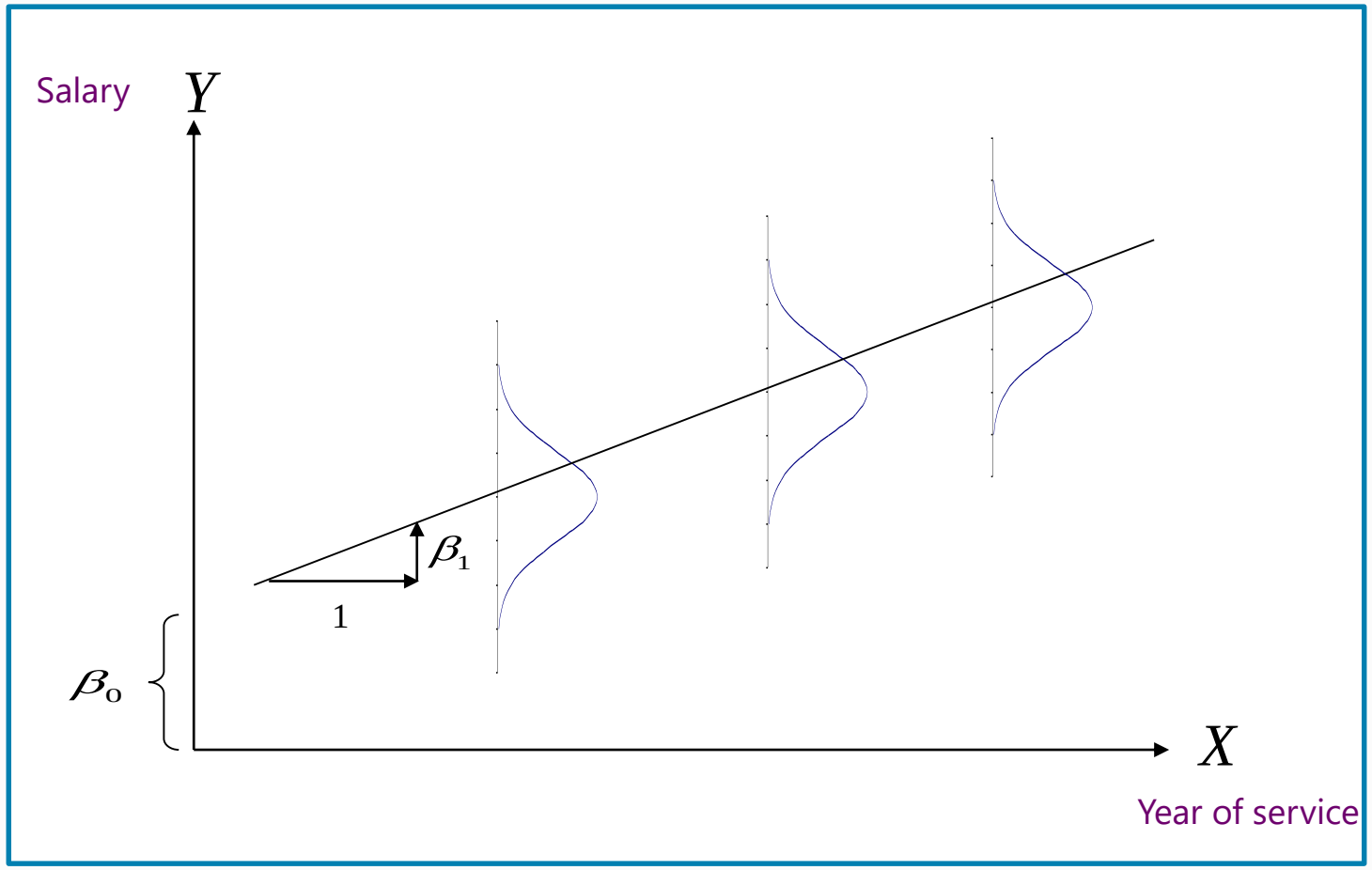
intercept

slope (coefficient for X)

Random error

Assume to have mean 0 and standard deviation “s”
Assume to follow a normal distribution
Assume to be independent

More details

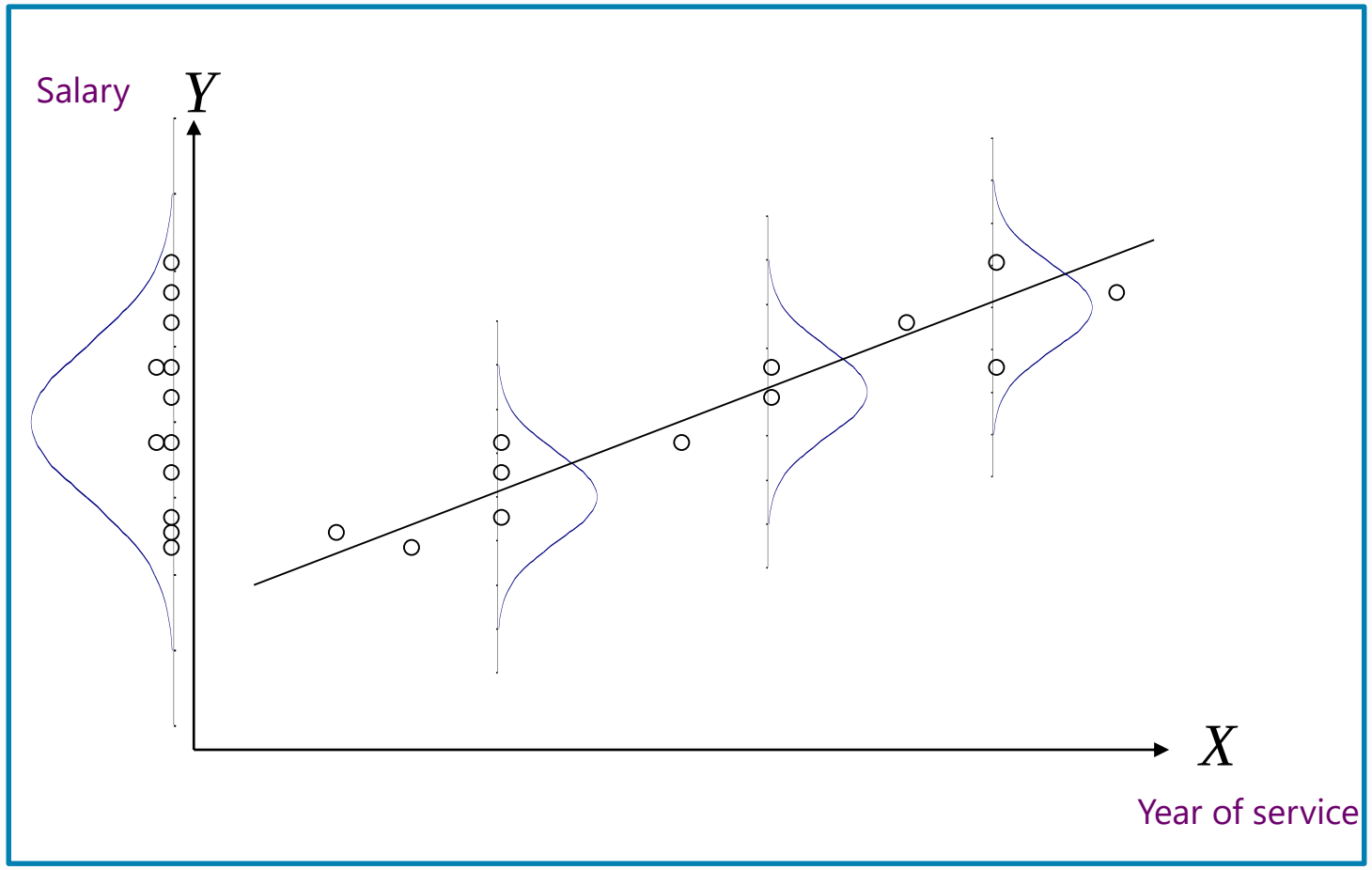


Slope = change in expected response for an unit increase in X

Intercept = expected response when X is 0

Random error is represented by the spread of the bell-shape curve about the model

More details



Slope = change in expected response for an unit increase in X

Intercept = expected response when X is 0

Random error is represented by the spread of the bell-shape curve about the model

Example 1 revisit

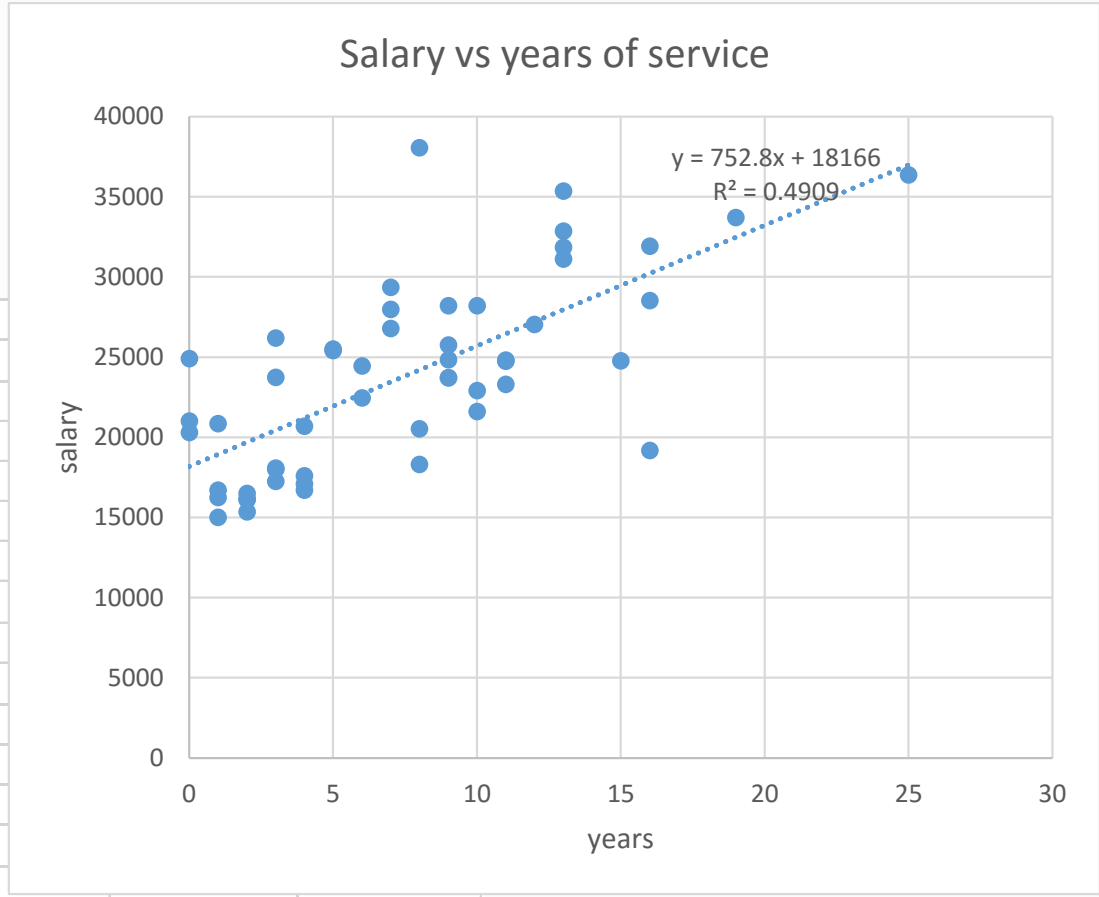
How should one assess the fit of the relationship?

Excel > Data > Data Analysis > Regression

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.700668985				
R Square	0.490937027				
Adjusted R Square	0.480755767				
Standard Error	4263.915925				
Observations	52				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	876680907	876680907	48.21967542	7.34138E-09
Residual	50	909048950.8	18180979.02		
Total	51	1785729858			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	18166.14755	1003.658162	18.09993506	1.34305E-23	16150.240
X Variable 1	752.7977574	108.4092089	6.944038841	7.34138E-09	535.05145

Model assessment

Parameter assessment



t-statistics = ratio of coef and SE(coef)
a measure of "signal-to-noise"
Large "signal-to-noise" leads to small "P-value",
which implies "statistical significance"

Model assessment

Measures of fit:

- R-square:
  - coefficient of determination
  - proportion of variation in the response attributable to fitting the model
  - between 0 and 1

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

numerator = squared deviation due to the model
denominator = squared deviation due to no model

- R-square adjusted:
  - R-square adjusted for the size / complexity of the model

$$R_a^{\ 2} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2/(n-p)}{\sum(Y_i - \bar{Y})^2/(n-1)}$$

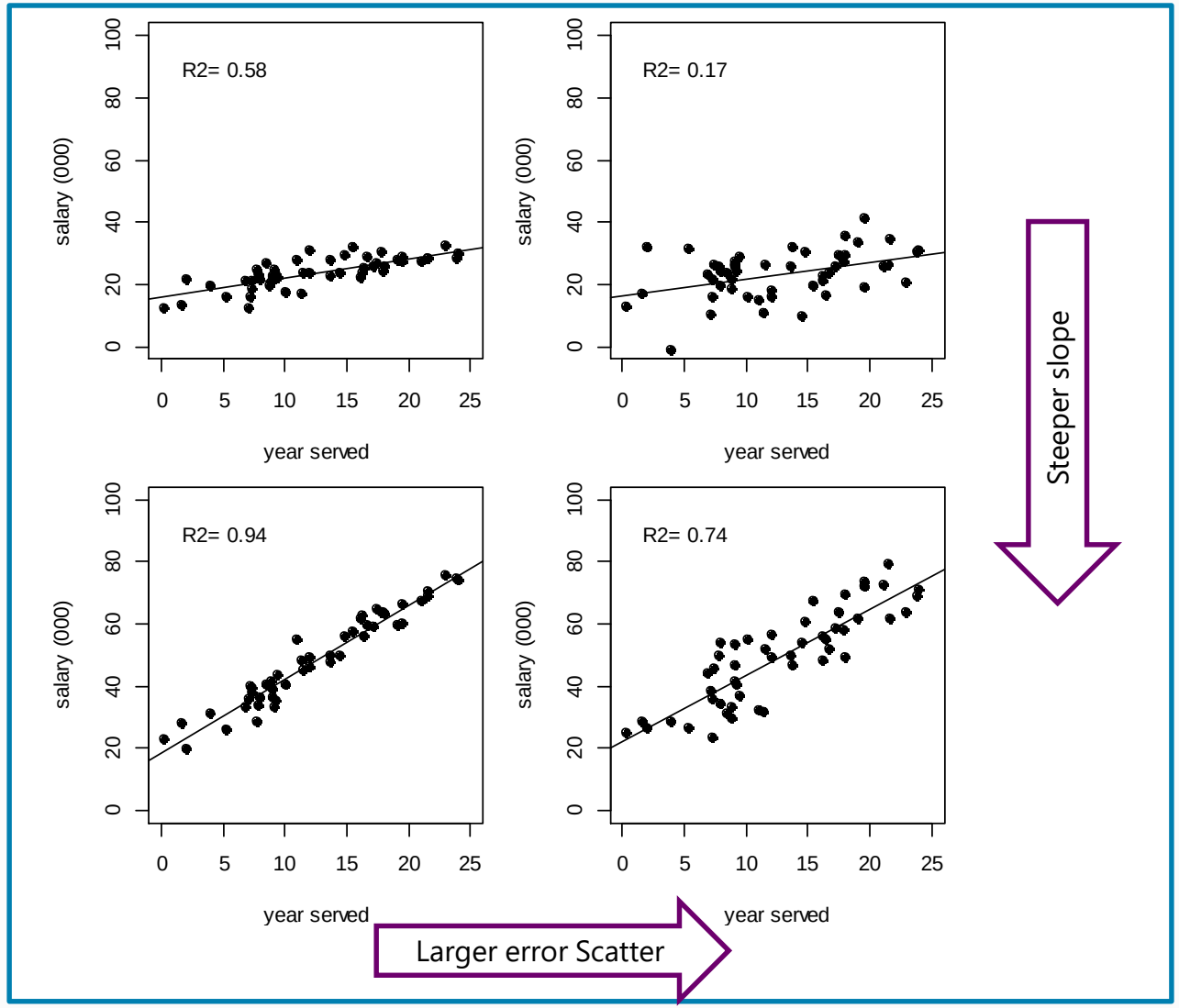
numerator of the quotient = mean squared error due to the model
denominator of the quotient = mean squared error due to no model
p = number of parameters including the intercept
n = number of observations

R-square increases with number of terms in the model but not necessary R-square adj.

Model assessment

Examples:

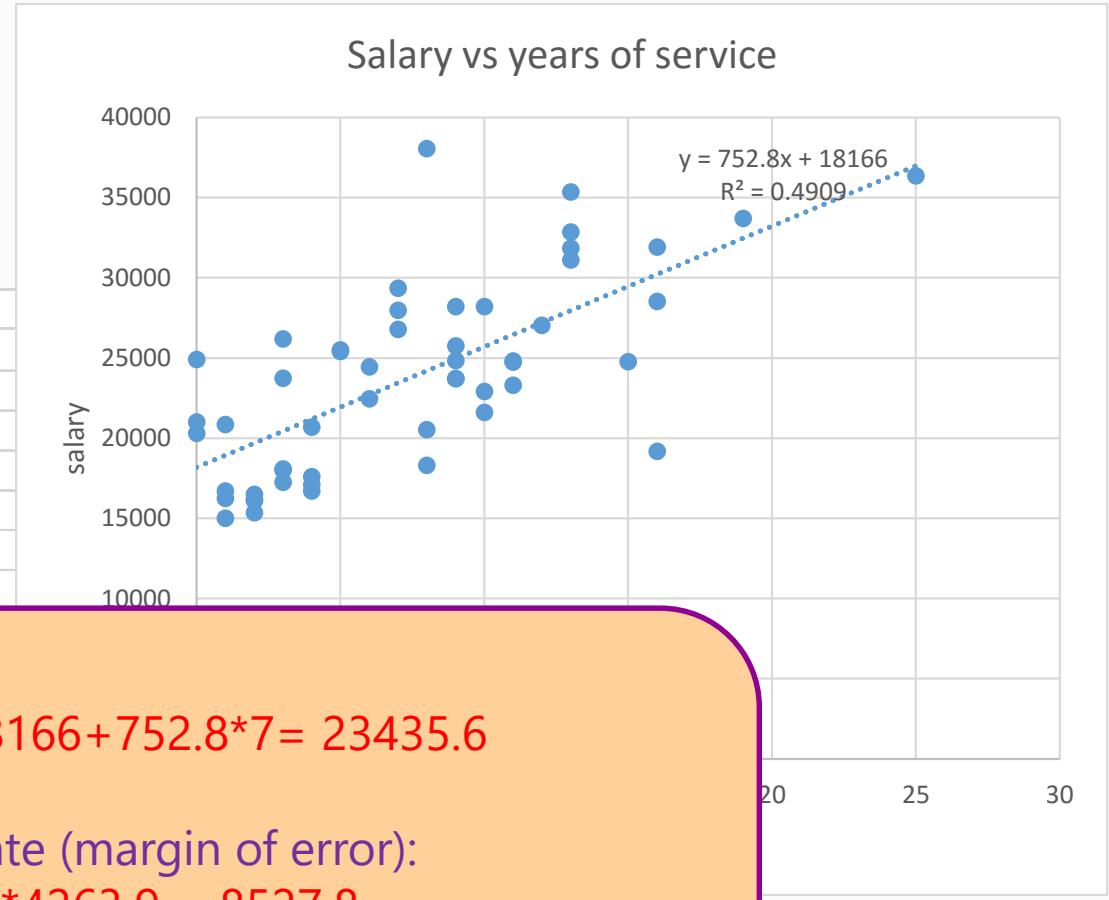
- Changes in R-sq due to
  - Varying slope (different rate of change in salary as years increases)
  - Varying noise variation (different level of scatterness)
- R-square reflects the “significance” of the linear relationship measured by both the line “steepness” and the data “scatterness.”
- High R-square is associated with high predictability.



Example 1 revisit

Estimate the salary range for a person with 7 years of experience.

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.700668985					
R Square	0.490937027					
Adjusted R Square	0.480755767					
Standard Error	4263.915925					
Observations	52					
ANOVA						
	df	*SS*				
Regression	1	876680907				
Residual	50	909048950.8				
Total	51	1785729858				
	Coefficients	*Standard Error*				
Intercept	18166.14755	1003.658162				
X Variable 1	752.7977574	108.4092089				



Estimate expected salary: 18166+752.8*7= 23435.6

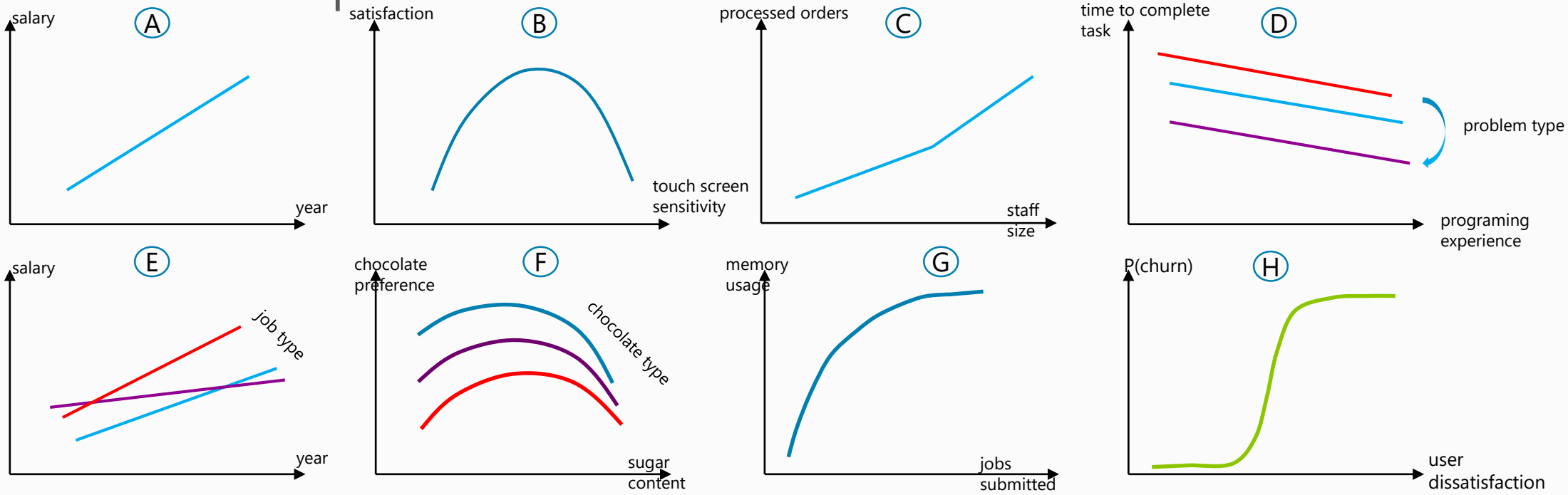
Uncertainty about this estimate (margin of error):
Roughly 2*standard error = 2*4263.9 = 8527.8

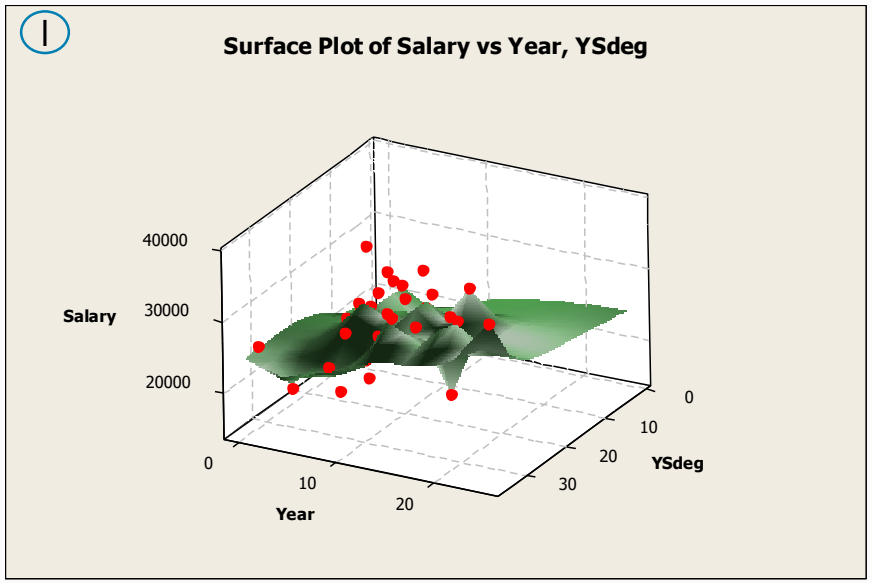
The expected salary range is estimated between 14908 and 31963

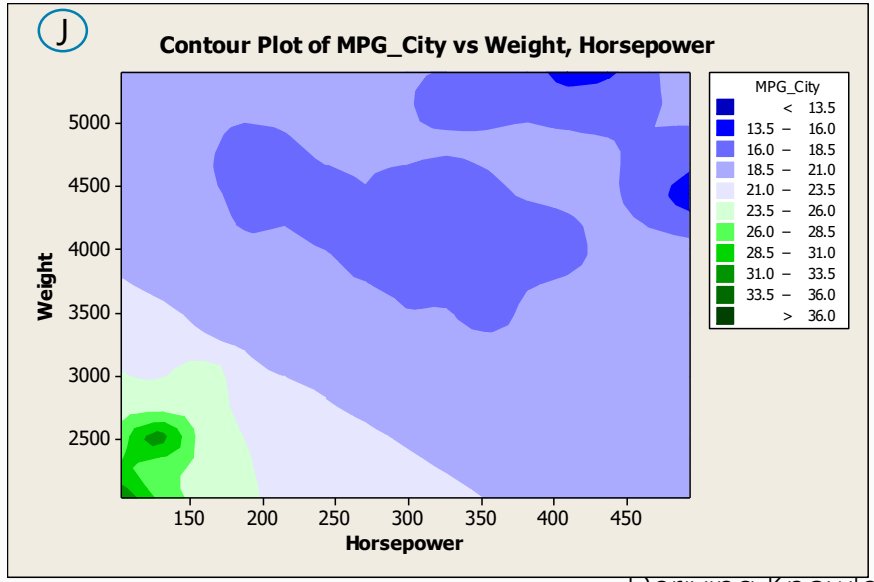
Beyond straight line

- How do we know we need more than a straight line?
- Data visualization at the raw data level
- Assess model fit through metrics such as R-square & adjusted R-square
- Data visualization after fitting a model to the data
- Multiple inputs? Plane or surface in higher dimension?

Other examples

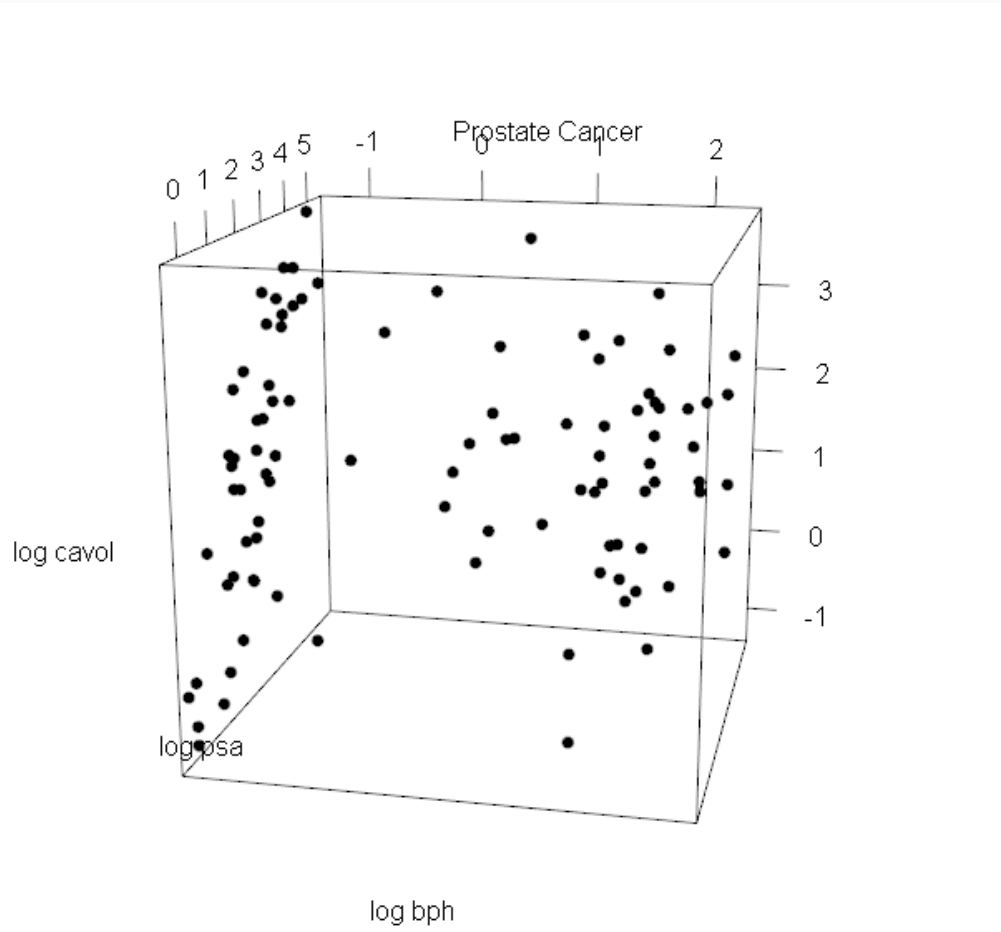






Rotate 3D

Visualize a 3D relationship interactively
Example:
Log Prostate tumor volume (log cavol)
Log Prostate specific antigen (log psa)
Log Benign prostatic hyoerplastia (log bph)



Graphical assessment of model

- Graphical approach:

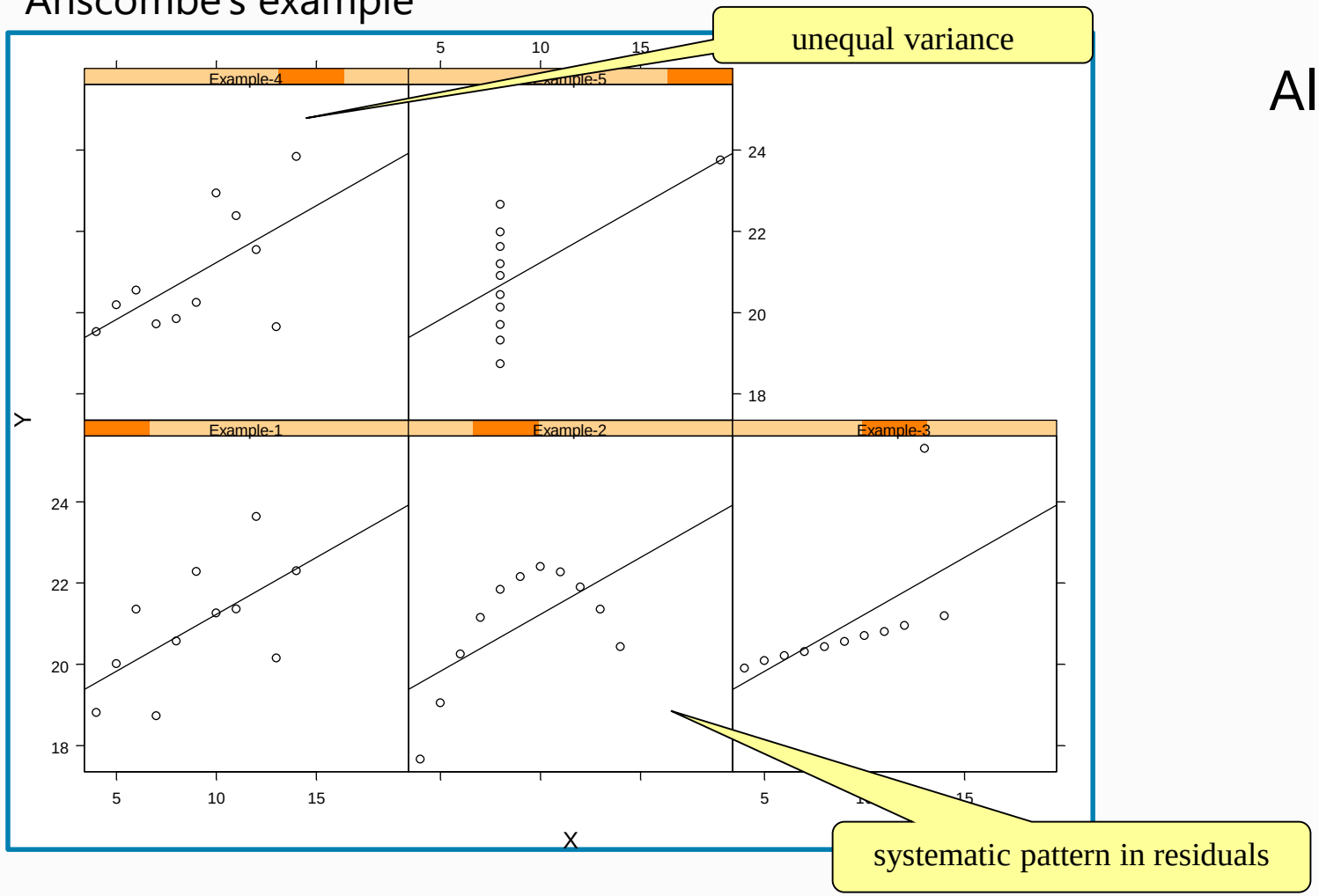
  Why bother with graphs when we have metrics? (see Anscombe's example)

  - Scatter plot of residual vs. fitted values
    - random and even scattering of residuals about a zero line
  - Scatter plot of residual vs. X variable
    - random and even scattering of residuals across X
  - Probability plot of residual
    - straight upward line suggesting conformance to specified distribution
  - Time plot of residual over time if possible
    - Random scattering over time (no time dependence)

- Difficult to visualize high dimensional data
  - Above plots can discover anomalies even in high dimensional data

Graphical assessment

Anscombe's example



All 5 examples have the same:

R-sq=0.3854

b0 = 18.43

b1 = 0.28

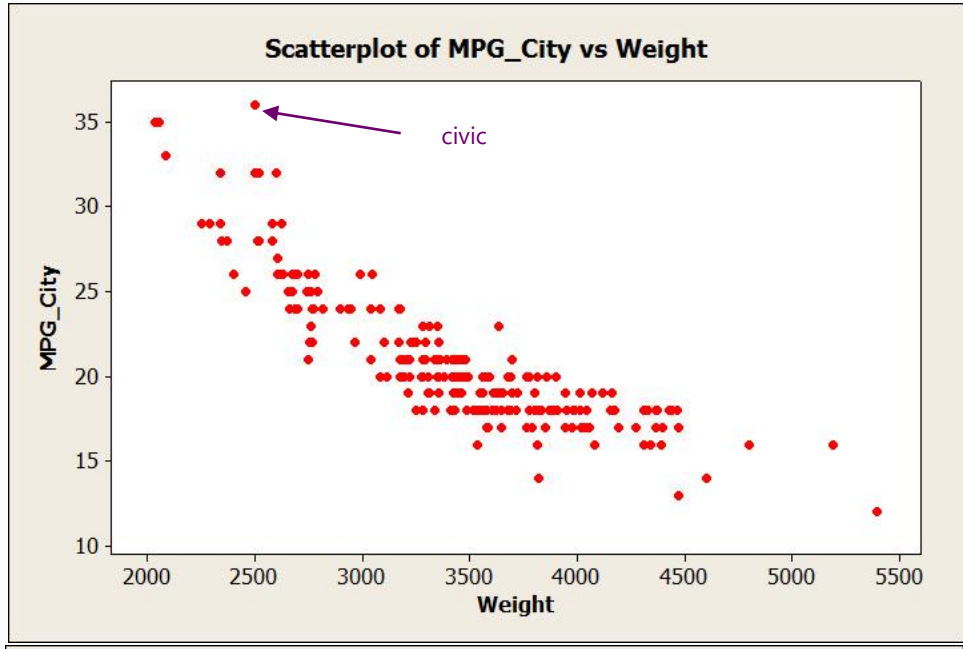
Example 2

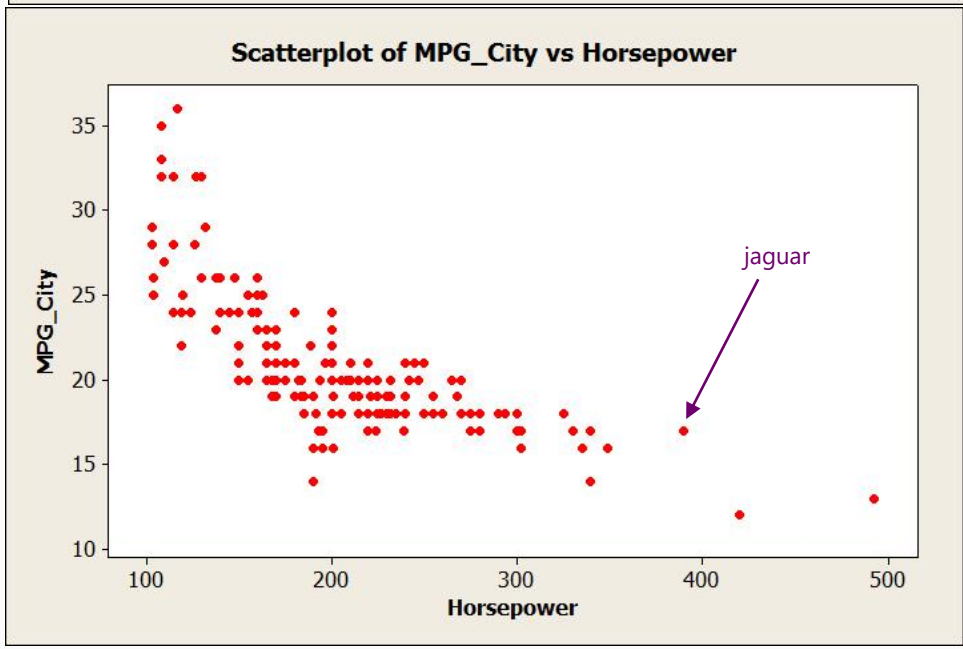
A vehicle's fuel economy is measured by miles per gallon (mpg). It varies depending on engine size, engine configuration, vehicle weight, types of vehicle, and many other factors. Given a sample of sedans, can one use these vehicle features to estimate fuel economy?

Is there a relationship between fuel economy and weight and horsepower in this data?

How would one estimate such relationship?

What does one use that estimated relationship?





Example 2 continue

If we were to find a relationship between fuel economy and weight and horsepower, we should be able to estimate the fuel economy of a particular vehicle using the model. The vehicle is the Acura TSX which weighs 3230 lbs and has 200 horsepower.

cars.csv - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW POWERPIVOT SAS

Normal Page Break Preview Page Layout Custom Views — Workbook Views
Ruler Formula Bar Gridlines Headings — Show
Zoom 100% Zoom to Selection — Zoom
New Window Arrange All Freeze Panes Split Hide Unhide View Side by Side Synchronous Scrolling Reset Window Position — Window
Switch Windows — Macros

AB2

Original data set

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Make	Model	Type	Origin	DriveTrain	MSRP	Invoice	EngineSiz	Cylinders	Horsepow	MPG_City	MPG_Highv	Weight	Wheelbas	Length
50	Buick	Century C	Sedan	USA	Front	$22,180	$20,351	3.1	6	175	20	30	3353	109	195
51	Buick	LeSabre C	Sedan	USA	Front	$26,470	$24,282	3.8	6	205	20	29	3567	112	200
52	Buick	Regal LS 4	Sedan	USA	Front	$24,895	$22,835	3.8	6	200	20	30	3461	109	196
53	Buick	Regal GS 4	Sedan	USA	Front	$28,345	$26,047	3.8	6	240	18	28	3536	109	196
54	Buick	LeSabre Li	Sedan	USA	Front	$32,245	$29,566	3.8	6	205	20	29	3591	112	200
55	Buick	Park Aven	Sedan	USA	Front	$35,545	$32,244	3.8	6	205	20	29	3778	114	207
56	Buick	Park Aven	Sedan	USA	Front	$40,720	$36,927	3.8	6	240	18	28	3909	114	207
57	Cadillac	Escalade	SUV	USA	Front	$52,795	$48,377	5.3	8	295	14	18	5367	116	199
58	Cadillac	SRX V8	SUV	USA	Front	$46,995	$43,523	4.6	8	320	16	21	4302	116	195
59	Cadillac	CTS VVT 4	Sedan	USA	Rear	$30,835	$28,575	3.6	6	255	18	25	3694	113	190
60	Cadillac	Deville 4d	Sedan	USA	Front	$45,445	$41,650	4.6	8	275	18	26	3984	115	207
61	Cadillac	Deville DT	Sedan	USA	Front	$50,595	$46,362	4.6	8	300	18	26	4044	115	207
62	Cadillac	Seville SLS	Sedan	USA	Front	$47,955	$43,841	4.6	8	275	18	26	3992	112	201
63	Cadillac	XLR conve	Sports	USA	Rear	$76,200	$70,546	4.6	8	320	17	25	3647	106	178
64	Cadillac	Escalade E	Truck	USA	All	$52,975	$48,541	6	8	345	13	17	5879	130	221
65	Chevrolet	Suburban	SUV	USA	Front	$42,735	$37,422	5.3	8	295	14	18	4947	130	219
66	Chevrolet	Tahoe LT	SUV	USA	All	$41,465	$36,287	5.3	8	295	14	18	5050	116	197
67	Chevrolet	TrailBlaze	SUV	USA	Front	$30,295	$27,479	4.2	6	275	16	21	4425	113	192
68	Chevrolet	Tracker	SUV	USA	Front	$20,255	$19,108	2.5	6	165	19	22	2866	98	163
69	Chevrolet	Aveo 4dr	Sedan	USA	Front	$11,690	$10,965	1.6	4	103	28	34	2370	98	167
70	Chevrolet	Aveo LS 4	Sedan	USA	Front	$12.585	$11.802	1.6	4	103	28	34	2348	98	153

Example 2 continue

Regression Analysis: Log10_MPG_City versus Weight, Horsepower

The regression equation is
Log10_MPG_City = 1.70 - 0.000093 Weight - 0.000342 Horsepower

Predictor	Coef	SE Coef	T	P
Constant	1.70040	0.01294	131.45	0.000
Weight	-0.00009305	0.00000551	-16.88	0.000
Horsepower	-0.00034244	0.00005094	-6.72	0.000

S = 0.0330888 R-Sq = 82.5% R-Sq(adj) = 82.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1.32111	0.66055	603.32	0.000
Residual Error	256	0.28029	0.00109		
Total	258	1.60140			

Source	DF	Seq SS
Weight	1	1.27162
Horsepower	1	0.04949

Regression model

Parameters significance

Model fit metrics

$$MPG = 10^{(b_0+b_1W+b_2H)}$$

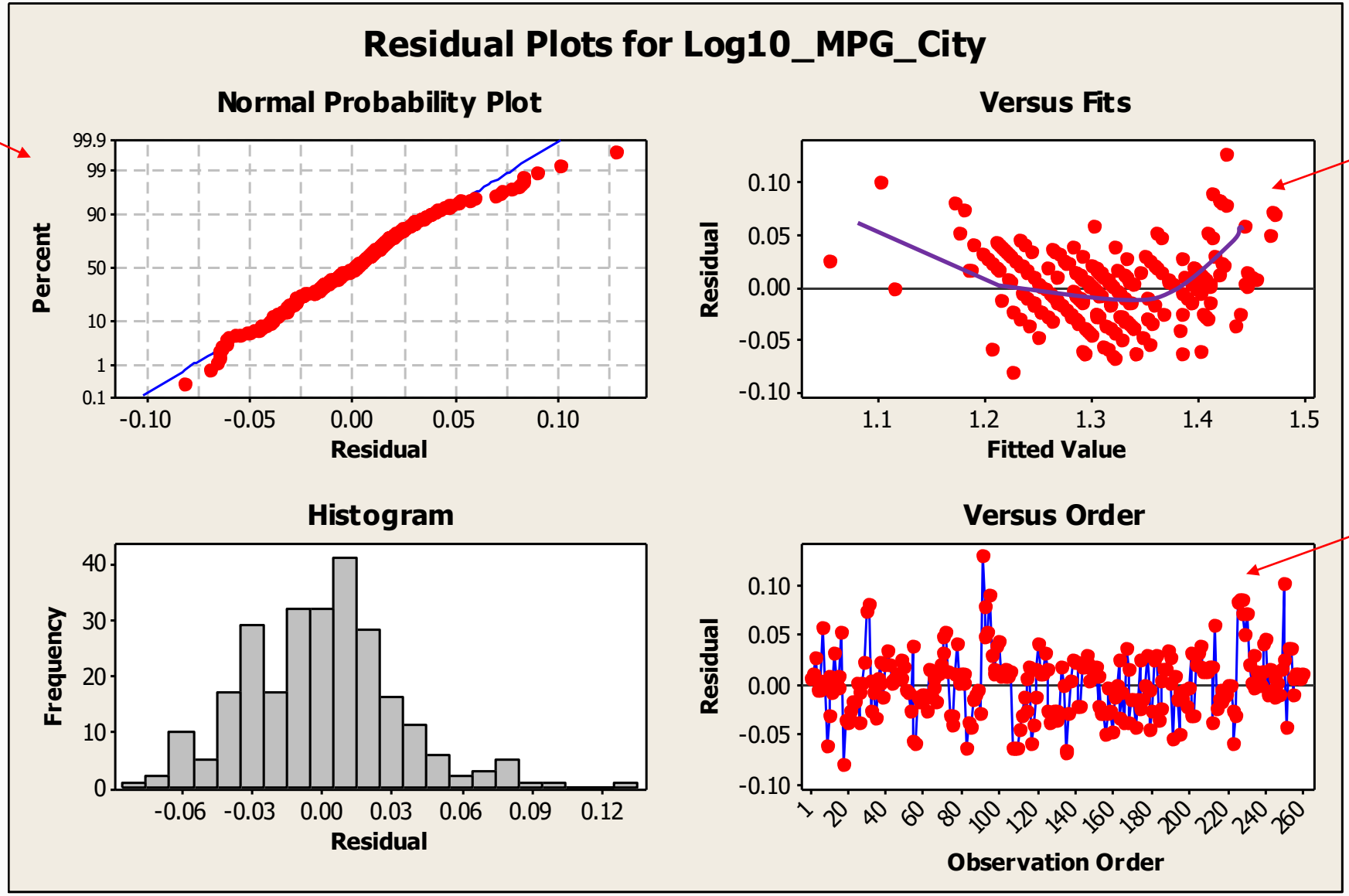
Parameter is 16.9 times of its own std error

Parameter is 6.7 times of its own std error

Example 2 continue

Ideal if data follows a straight line with little scatter

Ideal if bell shape



Ideal if even & random scattering about zero line

Some curvilinear pattern

Ideal if even & random scattering about zero line (meaningful only if observation order is meaningful)

Example 2 continue

Regression model

Parameters significance

Model fit metrics

$$MPG = 10^{(b_0 + b_1 W + b_{11} W^2 + b_2 H)}$$

Regression Analysis: Log10_MPG_City versus Weight, Horsepower, ...

The regression equation is

Log10_MPG_City = 2.05 - 0.000306 Weight - 0.000301 Horsepower
+ 0.000000 WeightSquare

Predictor	Coef	SE Coef	T	P
Constant	2.05270	0.04670	43.95	0.000
Weight	-0.00030564	0.00002774	-11.02	0.000
Horsepower	-0.00030084	0.00004618	-6.51	0.000
WeightSquare	0.00000003	0.00000000	7.79	0.000

S = 0.0297975 R-Sq = 85.9% R-Sq(adj) = 85.7%

Analysis of Variance

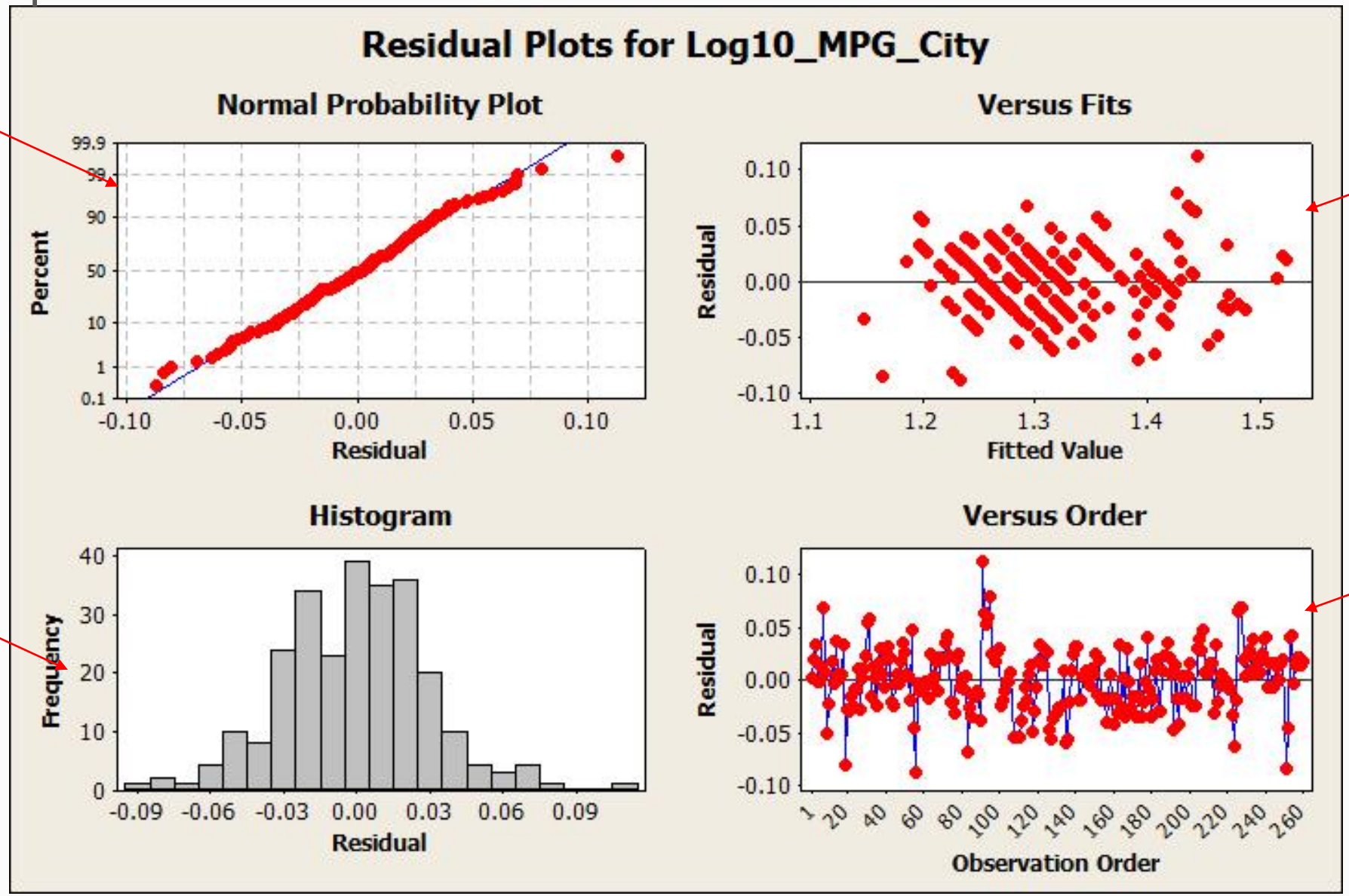
Source	DF	SS	MS	F	P
Regression	3	1.37498	0.45833	516.20	0.000
Residual Error	255	0.22641	0.00089		
Total	258	1.60140			

Source	DF	Seq SS
Weight	1	1.27162
Horsepower	1	0.04949
WeightSquare	1	0.05387

Example 2 continue

Ideal if data follows a straight line scatter

Ideal if bell shape



Ideal if even & random scattering about zero line

Ideal if even & random scattering about zero line (meaningful only if observation order is meaningful)

Example 2 continue

PLIM1= lower prediction limit
PLIM2= upper prediction limit

exercise2 ***

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
	MPG_City	Log10_MPG_City	MPG_Highway	Log10_MPG_Highway	Weight	WeightSquare	Horsepower				PLIM1	PLIM2	
1	24	1.38021	31	1.49136	2778	7717284	200	3230	200	10432900	1.26410	1.38181	
2	20	1.30103	28	1.44716	3575	12780625	270						
3	18	1.25527	24	1.38021	3880	15054400	225						
4	18	1.25527	24	1.38021	3893	15155449	225						
5	22	1.34242	31	1.49136	3252	10576504	170						

95% prediction limits for city fuel economy of Acura TSX are 18.4, 24.1 mpg

With high confidence, the Acura TSX city fuel economy is between 18.4 mpg and 24.1 mpg.

The published Acura TSX city fuel economy is **22** mpg.

Basic questions revisit

$$y = f(x) + \varepsilon$$

- Is there a relationship? y , x's
- What kind of relationship is it? $f(\,)$
- How strong is the relationship? ε
- Are there anomalies in the relationship? $y \sim f(x)$
- What is the scope of the relationship? *range of y , region of x's*
- What could be done from knowing this relationship?
- Should there be additional information to improve the accuracy of the relationship?

Model for fuel economy example

$$\log_{10} y = \beta_0 + \beta_1 w + \beta_{11} w^2 + \beta_2 h + \varepsilon$$

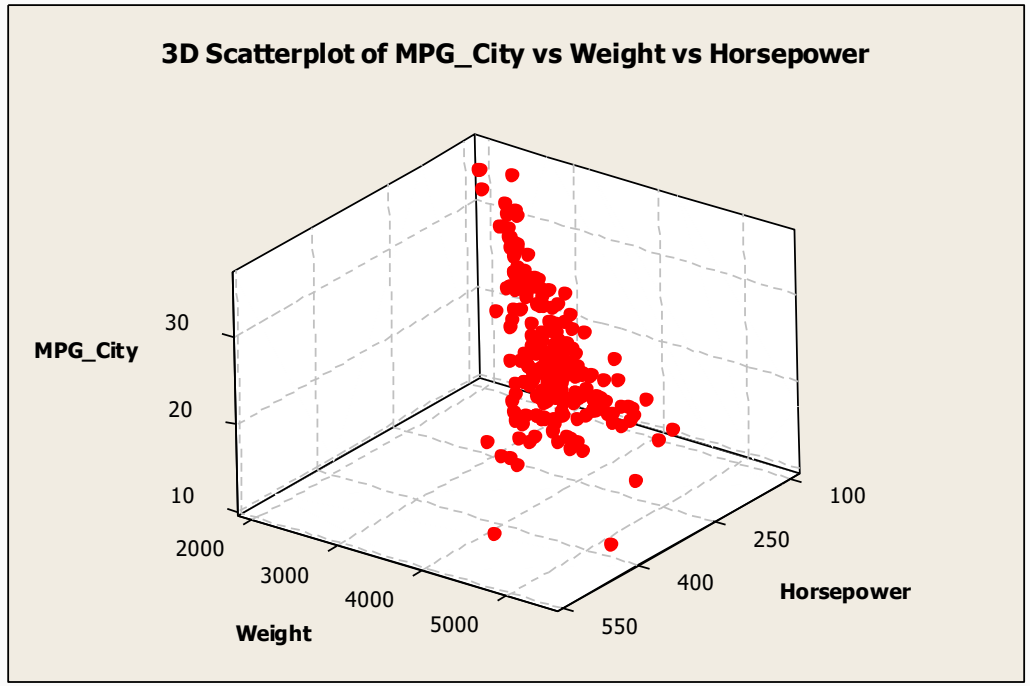
response

f(inputs)

noise, random error

Multiple linear regression

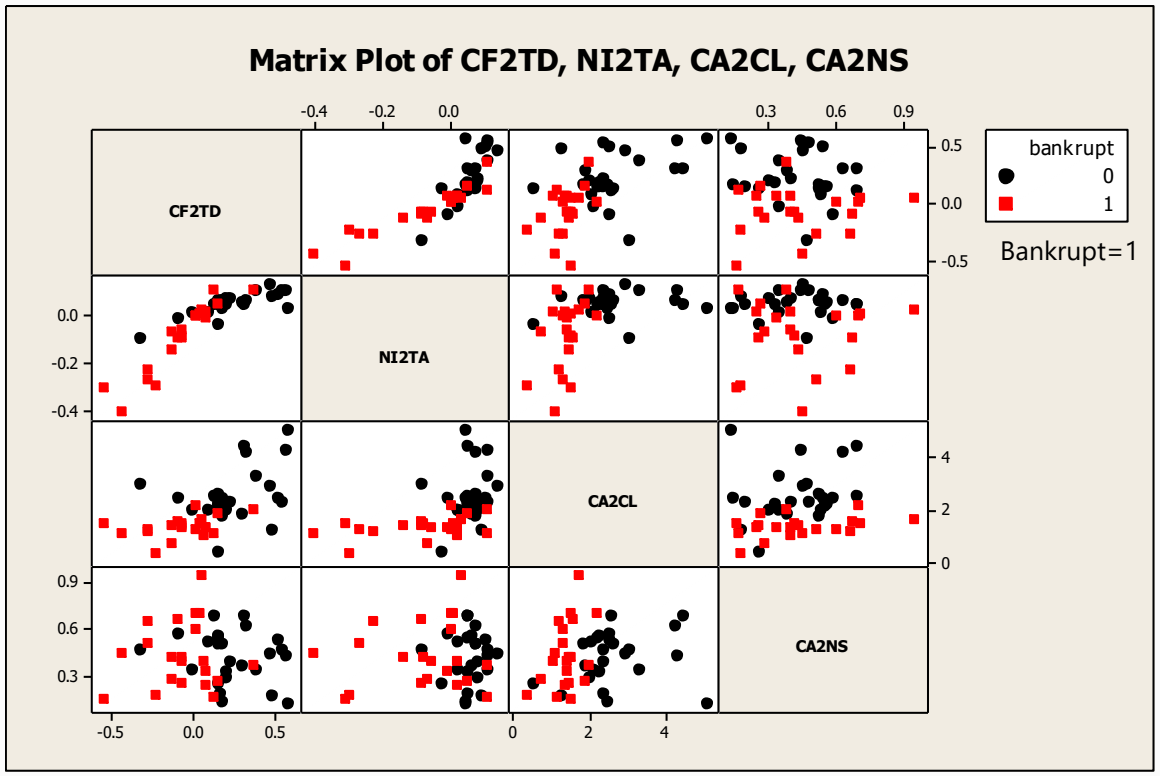
- Introduce and illustrate multiple linear regression
  - Model with 2 input variables
  - Model with 3 slopes (polynomial model)
  - Transformed response (stabilize scatter)



Logistic regression

In a Moody's report, a number of financial features of 21 bankrupted companies and 25 healthy companies were provided. 4 financial ratios were made available.

- Cash Flow to Total Debt (CF2TD),
- Net Income to Total Asset (NI2TA),
- Current Asset to Current Liability (CA2CL),
- Current Asset to Net Sales (CA2NS)



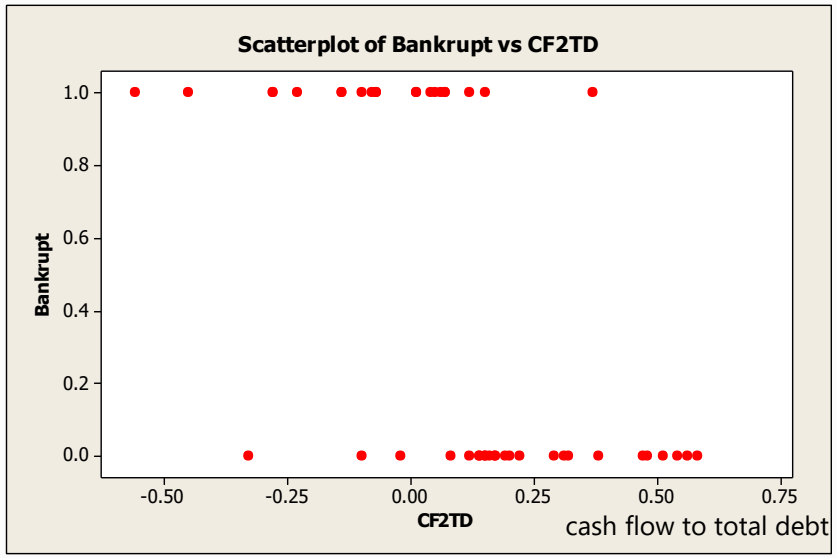
Which metric or metrics would be useful to predict propensity for "bankruptcy"? Can one use these metrics as leading indicator of financial struggle of a company?

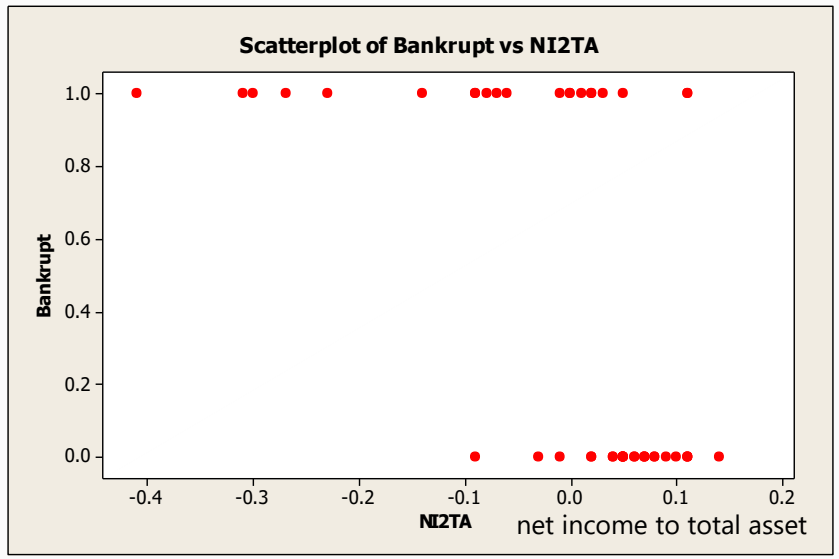
Is there a clear trend?

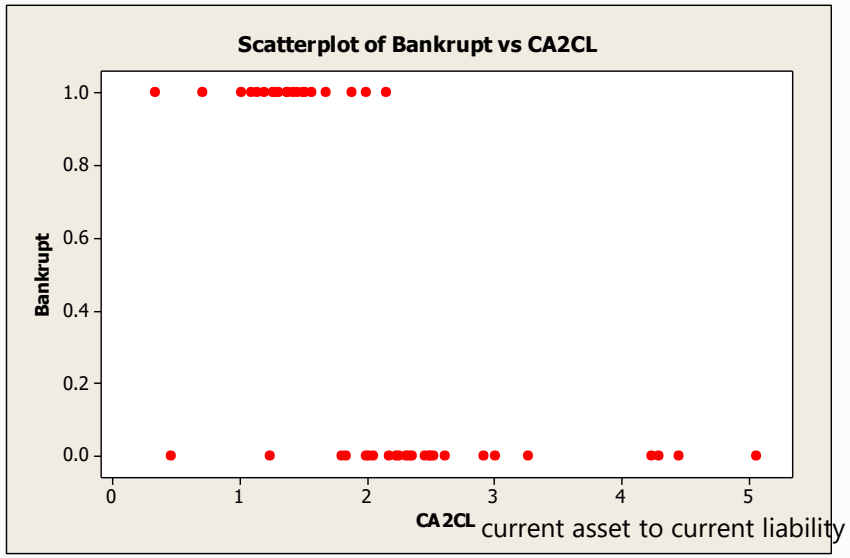
Is there a clear separation of bankruptcy among these financial metrics?

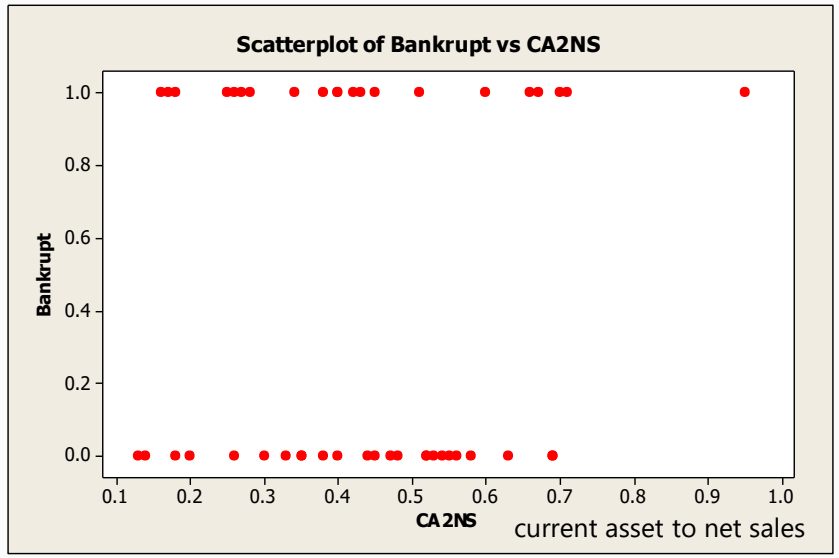
Does any combination of these financial metrics provide good prediction on bankruptcy?

Example 3









Are there obvious relationships between "bankruptcy" and these financial indicators?

CF2TD =cash flow to total debt
NI2TA = net income to total asset
CA2CL = current asset to current liability
CA2NS = current asset to net sales

Example 3 continue

The response is “categorical” in nature, namely, “bankrupt” or “not bankrupt”. Fitting a “straight line” through the data is not appropriate.

Logistic Regression is a general technique to fit a relationship between the probability of the categories and the input variables.

e.g. Single input X for a binary output Y. Suppose that Y=1 represents “bankrupt” and Y=0 represents “not bankrupt”

$$P(Y=1) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Probability of Y = 1

Relationship between P(Y=1) and Input x

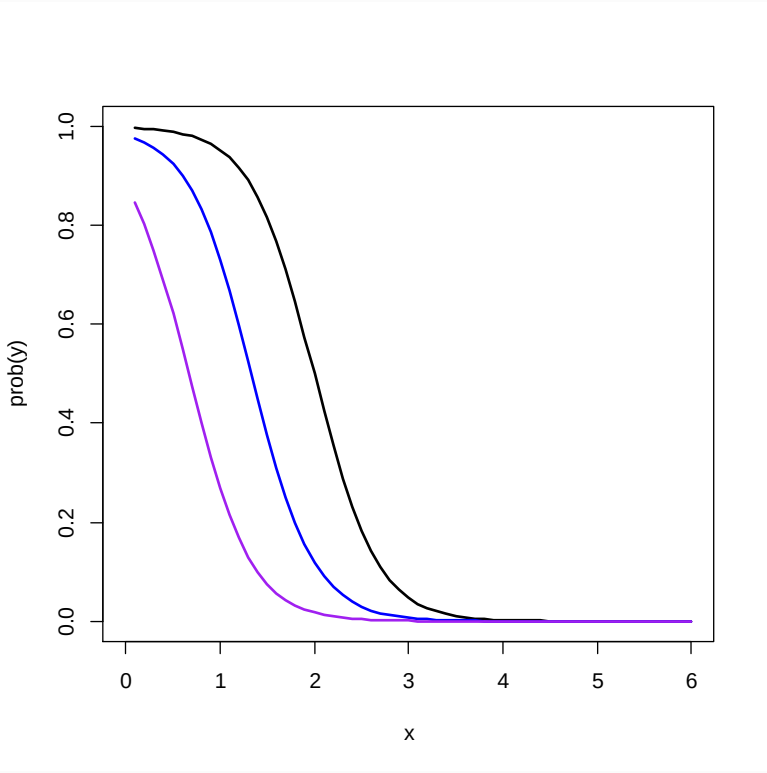
(link function)

Input x

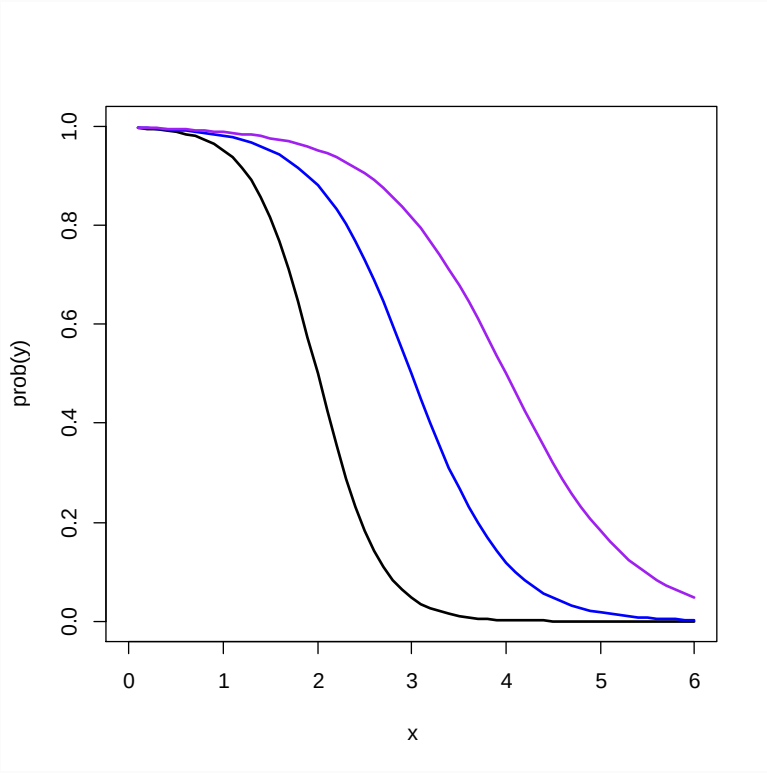
β_0, β_1 represent the parameters that determine the location of the “S” curve and the steepness of the “S” curve

Example 3 continue

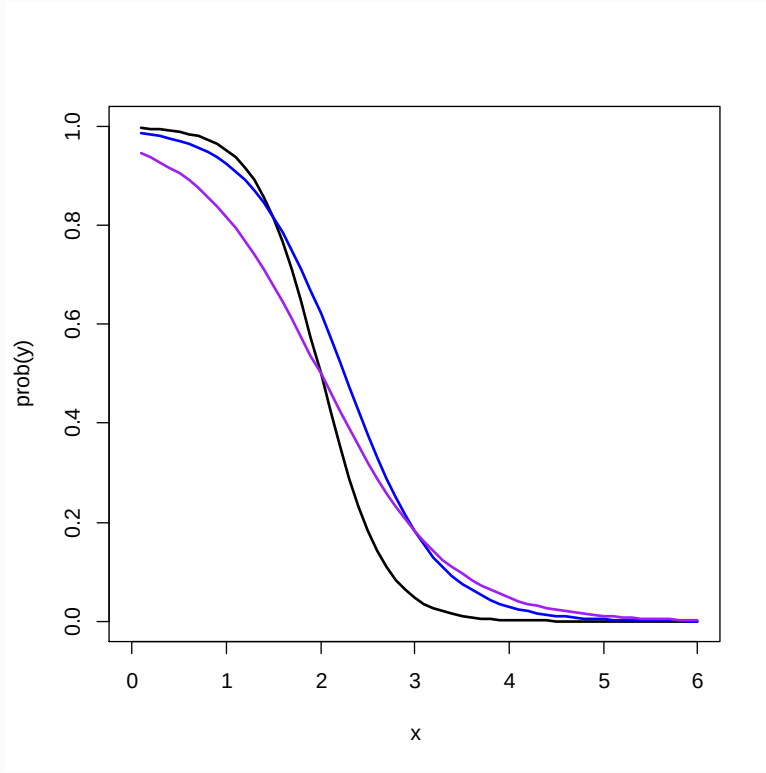
What changes the Logistic Regression model for “bankruptcy”?



Vary only β_0

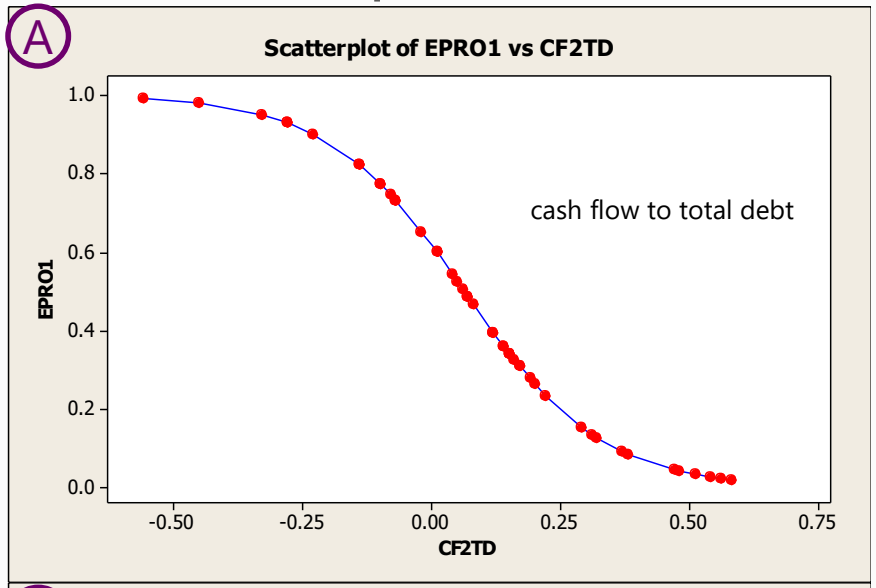


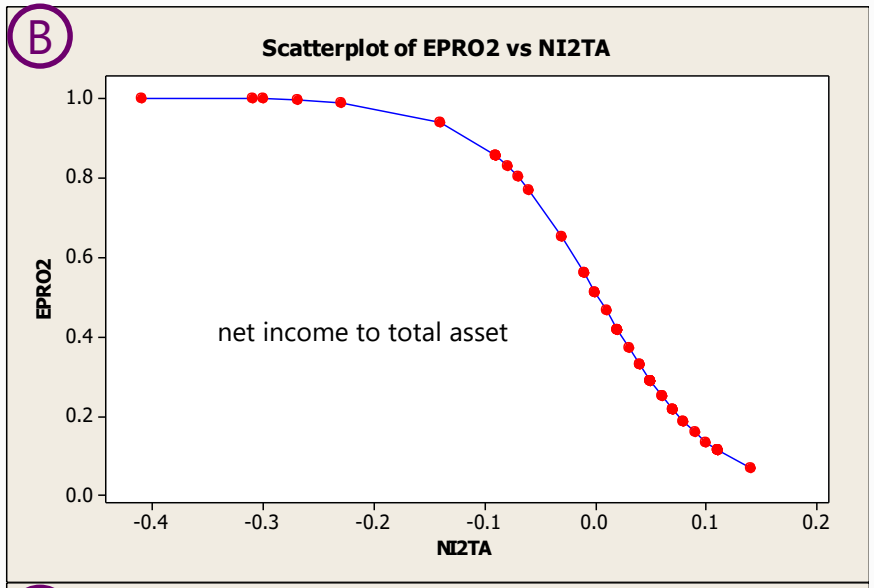
Vary only β_1

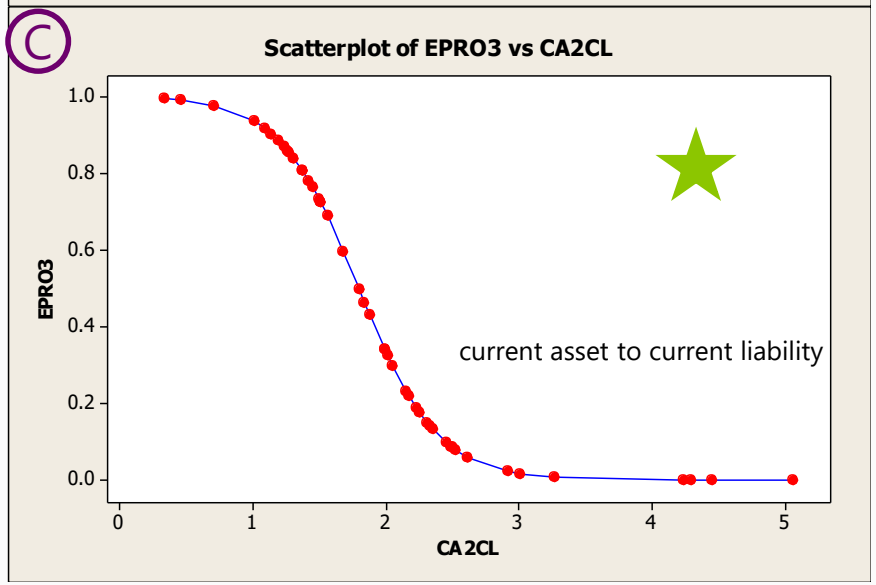


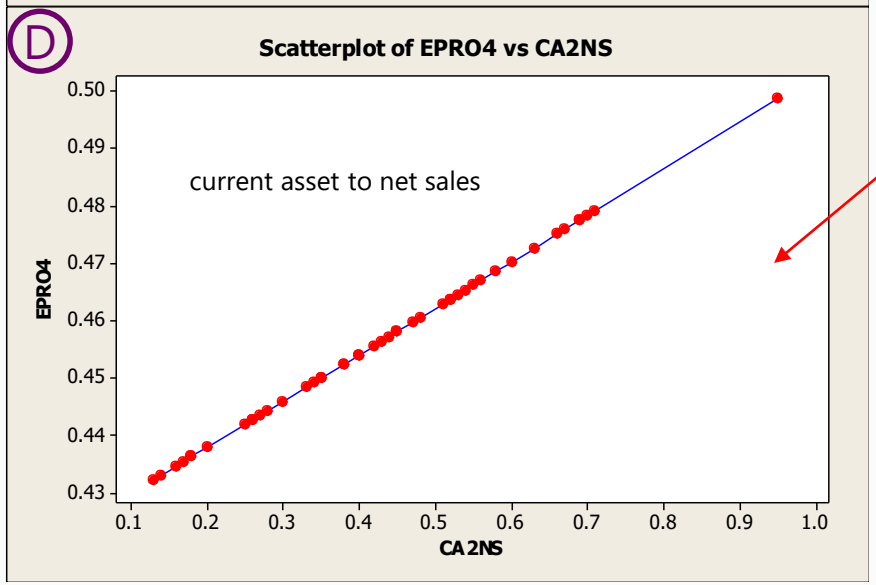
Vary both β_0, β_1

Example 3 continue









Which financial indicator best correlates with "bankruptcy"?

Note that this financial indicator has nearly no relationship with "bankruptcy". Note the vertical scale!

CF2TD =cash flow to total debt
NI2TA = net income to total asset
CA2CL = current asset to current liability
CA2NS = current asset to net sales

Example 3 continue

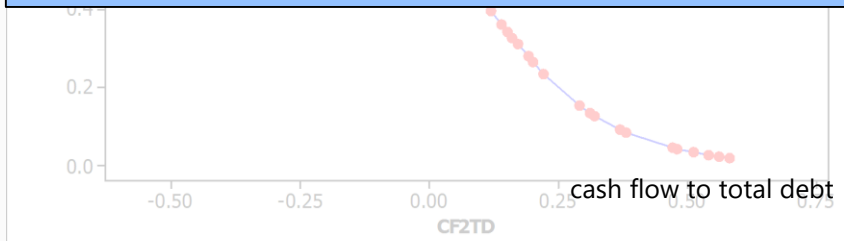
Scatterplot of EPRO1 vs CF2TD

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	0.479732	0.421095	1.14	0.255			
CF2TD	-7.54673	2.40776	-3.13	0.002	0.00	0.00	0.06

Log-Likelihood = -21.825

Test that all slopes are zero: G = 19.771, DF = 1, P-Value = 0.000



cash flow to total debt

Scatterplot of EPRO2 vs NI2TA

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	0.0492863	0.406763	0.12	0.904			
NI2TA	-19.1621	6.55740	-2.92	0.003	0.00	0.00	0.00

Log-Likelihood = -21.929

Test that all slopes are zero: G = 19.564, DF = 1, P-Value = 0.000

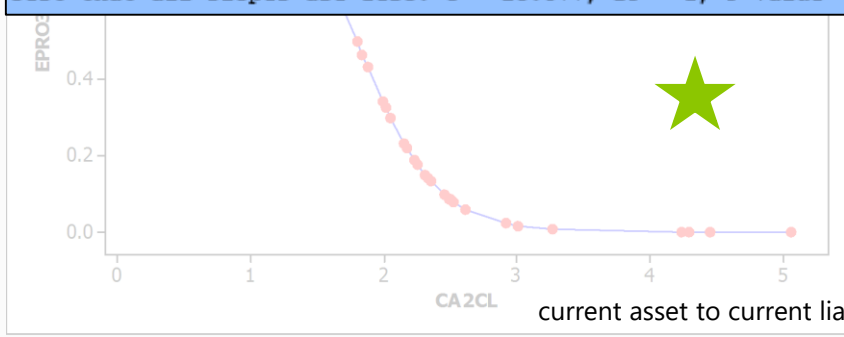
net income to total asset

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	6.05998	1.80991	3.35	0.001			
CA2CL	-3.37779	0.985388	-3.43	0.001	0.03	0.00	0.24

Log-Likelihood = -17.672

Test that all slopes are zero: G = 28.077, DF = 1, P-Value = 0.000



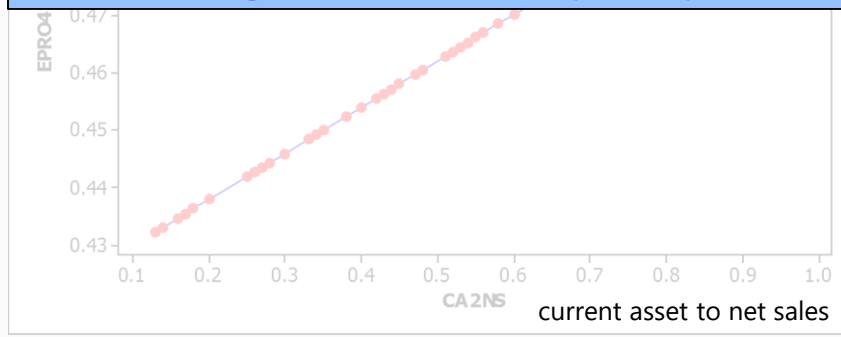
current asset to current liability

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-0.315285	0.763101	-0.41	0.679			
CA2NS	0.326086	1.62566	0.20	0.841	1.39	0.06	33.53

Log-Likelihood = -31.691

Test that all slopes are zero: G = 0.040, DF = 1, P-Value = 0.841



current asset to net sales

Which financial indicator best correlates with "bankruptcy"?

Based on the log-likelihood function, the best fit is the one with "CA2CL".

The one with "CA2NS" is not even statistically significant.

CF2TD =cash flow to total debt
NI2TA = net income to total asset
CA2CL = current asset to current liability
CA2NS = current asset to net sales

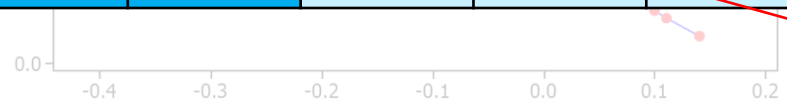
Example 3 continue

cash flow to total debt

CF2TD		Predict		
		Not	Bankrupt	
Observe	Not	22	3	25
	Bankrupt	5	16	21
		27	19	46

net income to total asset

NI2TA		Predict		
		Not	Bankrupt	
Observe	Not	22	3	25
	Bankrupt	7	14	21
		29	17	46



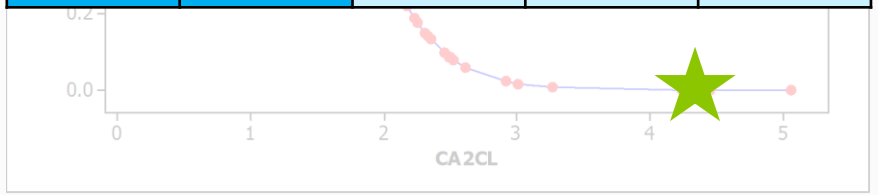
False Positive

False Negative

Which financial indicator best correlates with "bankruptcy"?

current asset to current liability

CA2CL		Predict		
		Not	Bankrupt	
Observe	Not	23	2	25
	Bankrupt	3	18	21
		26	20	46



current asset to net sales

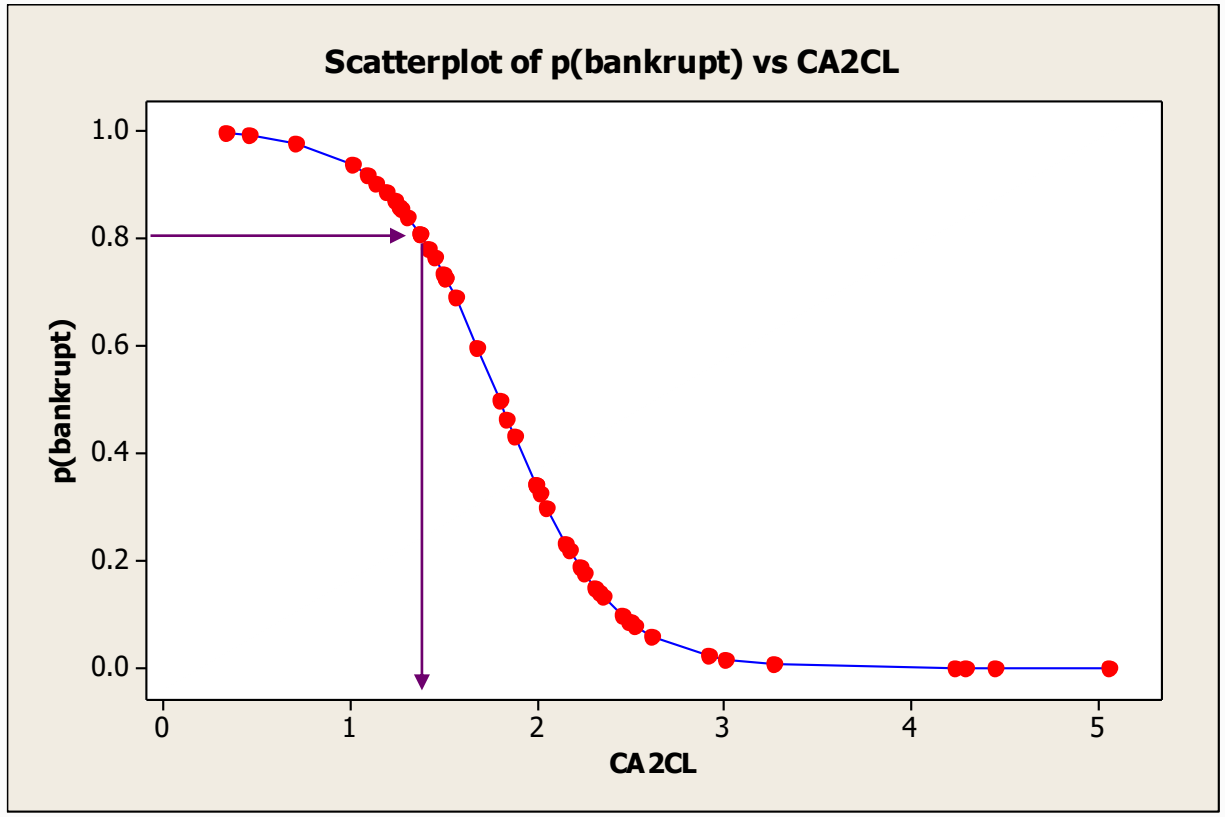
CA2NS		Predict		
		Not	Bankrupt	
Observe	Not	25	0	25
	Bankrupt	21	0	21
		46	0	46

Based on the error rate in these "confusion" tables, using CA/CL ratio would have a lower risk.

CF2TD =cash flow to total debt
NI2TA = net income to total asset
CA2CL = current asset to current liability
CA2NS = current asset to net sales

Example 3 continue

Conduct a logistic regression fit to all 4 financial indicators.
Check for statistically significant contributor(s) and fit metrics.
The financial indicator CA2CL is the best contributor.



The probability of bankruptcy decreases as CA/CL increases.

This model enables users to estimate / predict the chance of bankruptcy with a range of current asset to current liability. (CA/CL)

Another way to gage a company's health is to use the relationship in a reverse manner. For example, set a 80% threshold on P(bankruptcy) and flag a company near imminent bankruptcy when it has CA less than 1.3*CL.

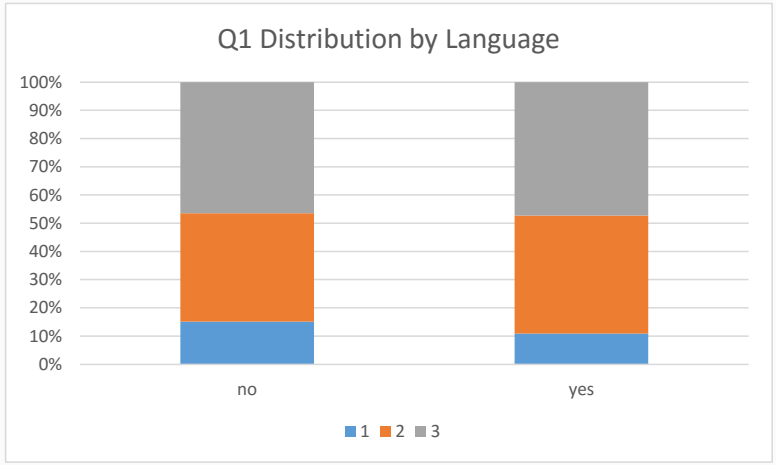
Other regression models

	Method	Input	Output	Other tools
Example 1	Linear regression	Single continuous	Single continuous	
Example 2	Multiple regression (including polynomial)	Multiple continuous and categorical	Single continuous	Regression tree; generalized additive model
Example 3	Logistic regression	Multiple continuous and categorical	Single binary	Classification tree; log linear model
Example 4	Multinomial regression	Multiple continuous and categorical	Single multinomial	Classification tree; log linear models

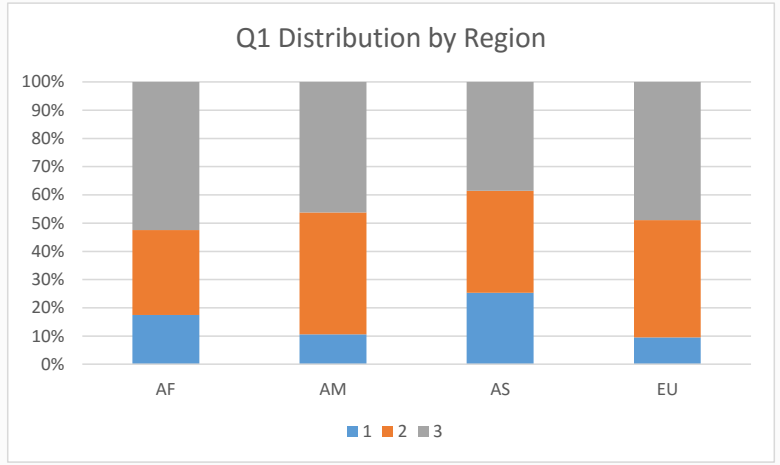
Multinomial regression

A software to process transaction was changed from version A to version B. The user experience from partners was measured through a survey. The overall satisfaction on version A and version B, satisfaction on version B specific functionality, and other background information were gathered. The product team wished to

a) figure out which attribute(s) determine the satisfaction on version B,
b) improve the user experience by acting on these attributes



Is there a relationship between Q1 and language?



Is there a relationship between Q1 and geographical regions?

Q1 = overall sat. with version B (3 points scale)

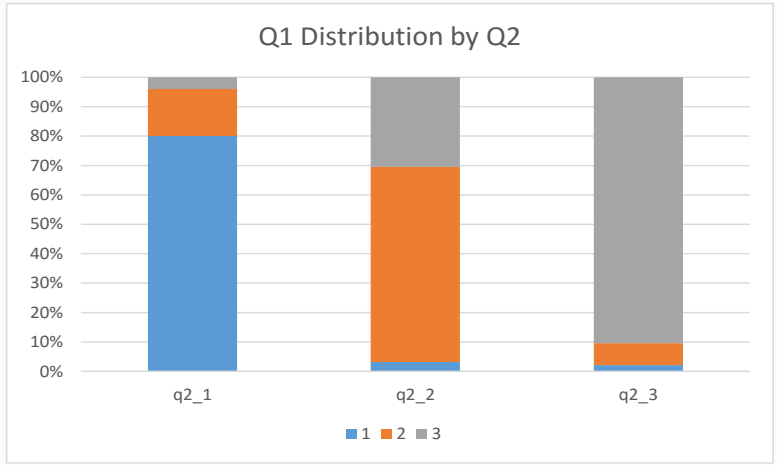
Language = yes for English, =no for non-English

Region = AF for Africa & Middle East, = AM for all Americas, = AS for Asia Pacific, = EU for Europe

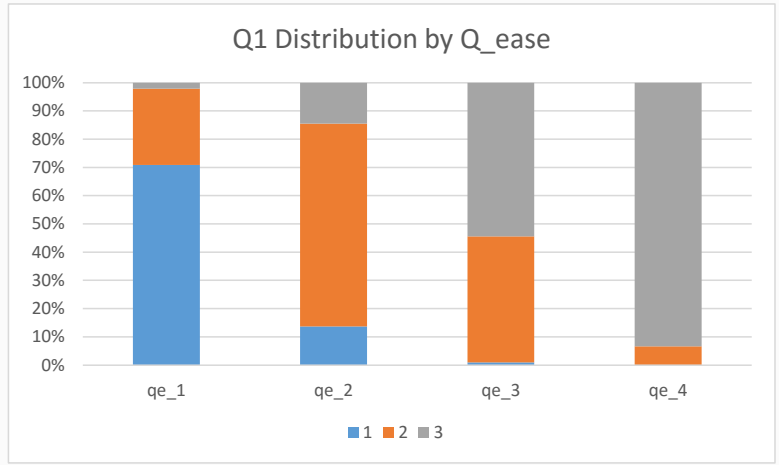
Example 4

A software to process transaction was changed from version A to version B. The user experience from partners was measured through a survey. The overall satisfaction on version A and version B, satisfaction on version B specific functionality, and other background information were gathered. The product team wished to

a) figure out which attribute(s) determine the satisfaction on version B,
b) improve the user experience by acting on these attributes



Is there a relationship between Q1 and Q2?



Is there a relationship between Q1 and ease of use?

Q1 = overall sat. with version B (3 points scale)

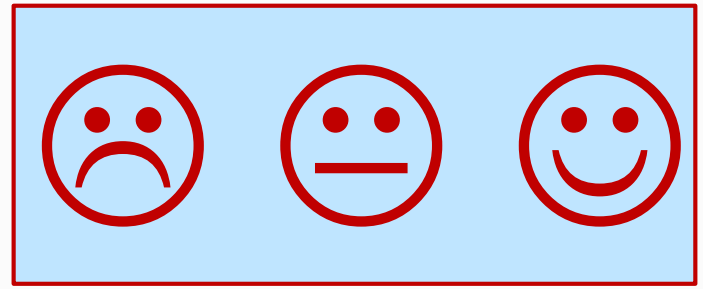
Q2 = overall sat. with version A (3 points scale)

Q_ease = sat. with version B ease of use (4 points scale)

Example 4 continue

One Multinomial model:

- More than 2 categories! (beyond binary)
- Assume **ordinal** rating
- Assume 3 ratings only



p_1 Probability of rating 1

p_2 Probability of rating 2

p_3 Probability of rating 3

Given a survey question, these ratings are mutually exclusive. It is certain that the rating is one of these 3.

$$p_1 + p_2 + p_3 = 1$$

Since ratings are "ordinal" in nature, namely, a rating of 1 is considered less desirable than a rating of 2 and so on, there is a relative ranking among these ratings. The actual distance between the ratings are "subjective" and difficult to quantify.

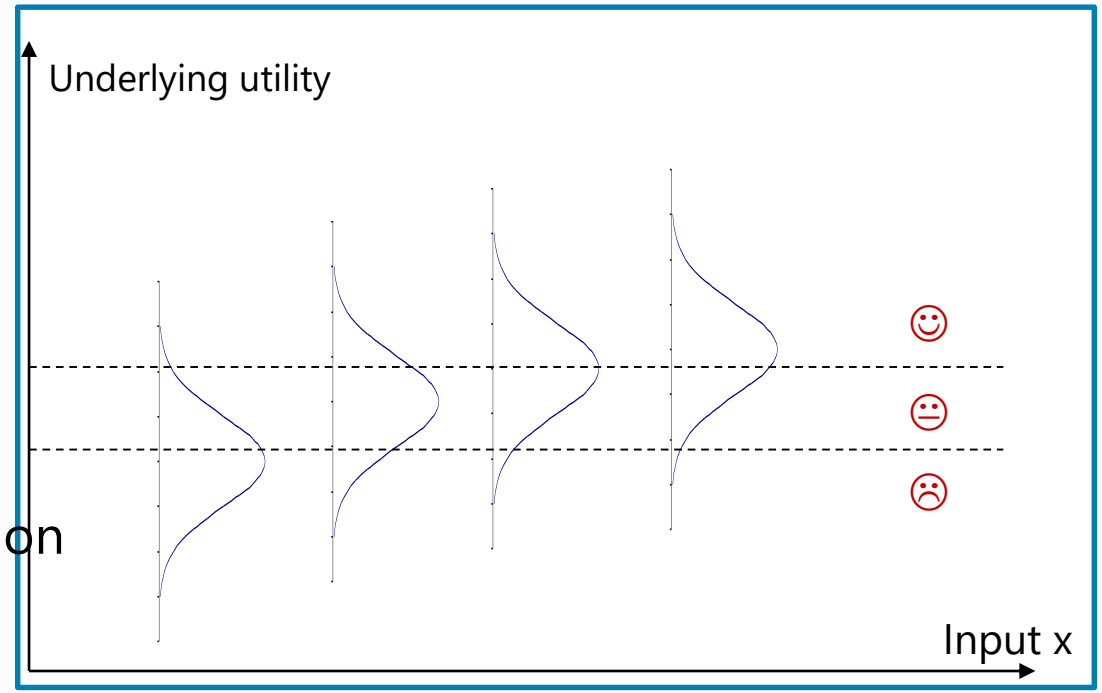
Example 4 continue

One Multinomial model:

- More than 2 categories! (beyond binary)
- Assume **ordinal** rating
- Assume 3 ratings only

Underlying utility?
Psychometric?
Unobservable metric?

e.g. y=Usability, x= hours of training
e.g. y=Taste test, x= amount of sugar
e.g. y=Post surgery mobility, x= pain mediation
e.g. y=Comfort level, x= leg room
…



The next 3 slides illustrate the "Cumulative Logit Model" capturing the ordinal nature of these 3 ratings and its probability increases with input x.

Example 4 continue

Let us focus on two particular questions.

		Question 2			
		1	2	3	
Question 1	1 ☹	40	4	2	46
	2 😐	8	83	7	98
	3 ☺	2	38	85	125
		50	125	94	269

Treat ratings from Q2 as input x.
The probability of Q1 rating increases with x.

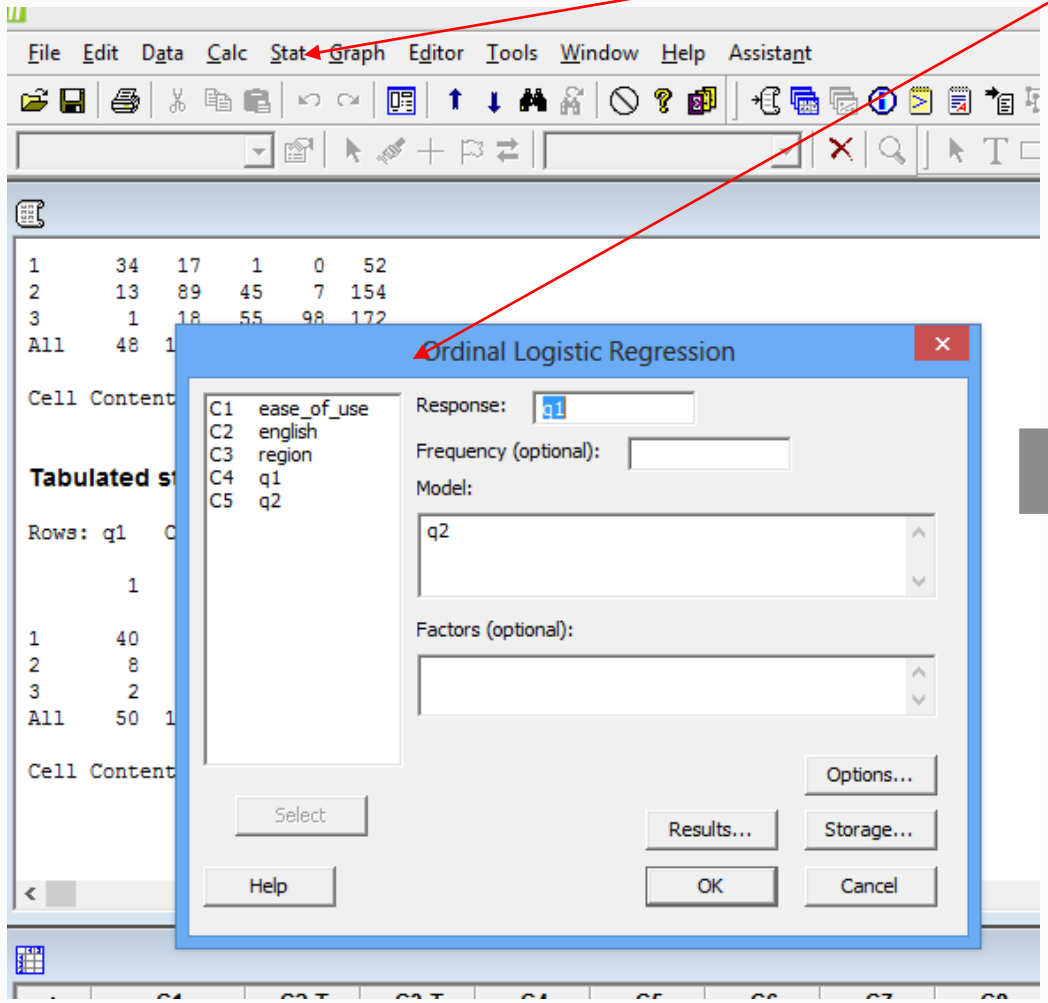
		Ease of use				
		1	2	3	4	
Question 1	1	34	17	1	0	52
	2	13	89	45	7	154
	3	1	18	55	98	172
		48	124	101	105	378

Treat ratings from "ease of use" as input x.
The probability of Q1 rating increases with x.

Example 4 continue

Fit a Multinomial model with MINITAB

Select “Regression”
Select sub menu “Ordinal Logistic Regression”



Ordinal Logistic Regression: q1 versus q2

Link Function: Logit

Response Information

Variable	Value	Count
q1	1	46
	2	98
	3	125
	Total	269

* NOTE * 269 cases were used
* NOTE * 109 cases contained missing values

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Const(1)	4.52693	0.524804	8.63	0.000			
Const(2)	7.85690	0.674089	11.66	0.000			
q2	-3.48953	0.306994	-11.37	0.000	0.03	0.02	0.06

Log-Likelihood = -164.412
Test that all slopes are zero: G = 223.163, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

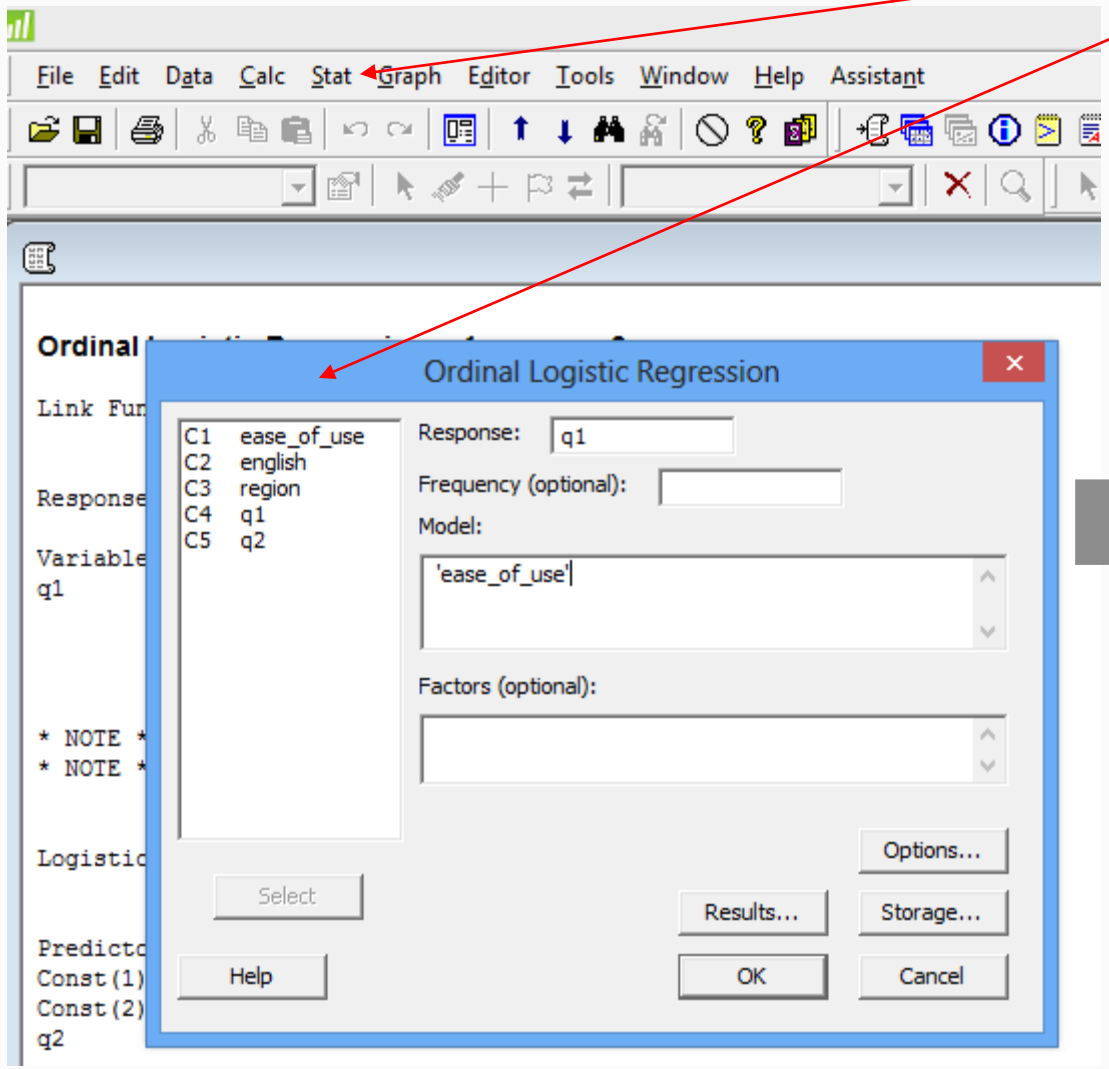
Method	Chi-Square	DF	P
Pearson	21.3738	3	0.000
Deviance	13.8988	3	0.003

Statistically significant coefficients because of the high signal-to-noise value in the Z statistics

Example 4 continue

Fit a Multinomial model with MINITAB

Select “Regression”
Select sub menu “Ordinal Logistic Regression”



Ordinal Logistic Regression: q1 versus ease_of_use

Link Function: Logit

Response Information

Variable	Value	Count
q1	1	52
	2	154
	3	172
	Total	378

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Const(1)	2.95702	0.369331	8.01	0.000			
Const(2)	6.66284	0.525867	12.67	0.000			
ease_of_use	-2.32171	0.180658	-12.85	0.000	0.10	0.07	0.14

Log-Likelihood = -232.466
Test that all slopes are zero: G = 288.799, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	3.22797	5	0.665
Deviance	3.42499	5	0.635

Statistically significant coefficients because of the high signal-to-noise value in the Z statistics

Example 4 continue

Choose a final model using fit metrics

Q1 as a function of Q2

```
Log-Likelihood = -164.412
Test that all slopes are zero: G = 223.163, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests
Method     Chi-Square  DF      P
Pearson       21.3738   3  0.000
Deviance      13.8988   3  0.003
```

Statistically significant => difference between observed and expected counts from the model is beyond random error

obs	q2_1	q2_2	q2_3
1	40	4	2
2	8	83	7
3	2	38	85

model	q2_1	q2_2	q2_3
1	37	10	0
2	12	78	6
3	1	37	88

Pearson statistics (Measure of model deviation): 21.3738

Q1 as a function of Q_ease

```
Log-Likelihood = -232.466
Test that all slopes are zero: G = 288.799, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests
Method     Chi-Square  DF      P
Pearson       3.22797   5  0.665
Deviance      3.42499   5  0.635
```

Statistically insignificant => small difference between observed and expected counts from the model

obs	qe_1	qe_2	qe_3	qe_4
1	34	17	1	0
2	13	89	45	7
3	1	18	55	98

model	qe_1	qe_2	qe_3	qe_4
1	31	19	2	0
2	16	90	41	7
3	1	15	58	98

Pearson statistics: Measure of model deviation: 3.2280

Example 4 continue

A software to process transaction was changed from version A to version B. The user experience from partners was measured through a survey. The overall satisfaction on version A and version B, satisfaction on version B specific functionality, and other background information were gathered. The product team wished to
a) figure out which attribute(s) determine the satisfaction on version B,
b) improve the user experience by acting on these attributes

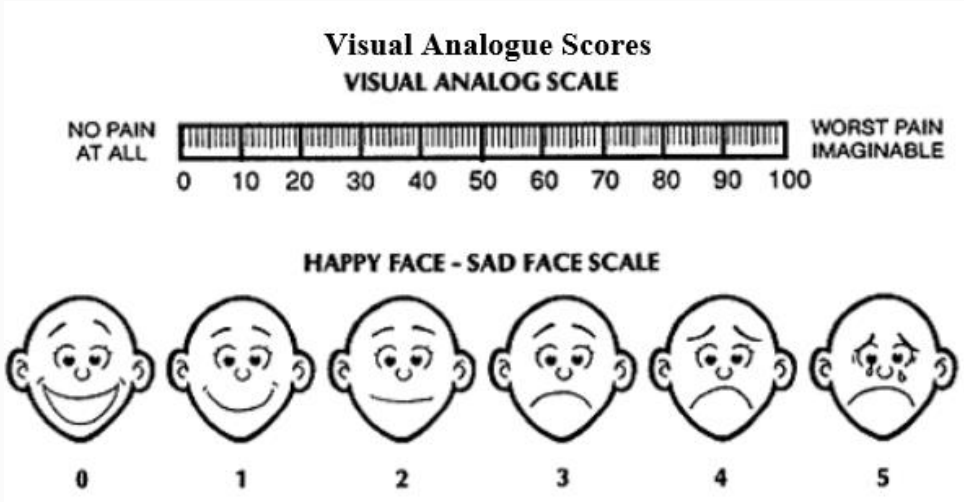
Among different attributes that influence "overall satisfaction of Version B", "ease-of-use" is the strongest contributing factor.

"Overall satisfaction of version A" is also important but not as strong, so are other attributes.

Improve "ease-of-use" would likely lead to overall satisfaction. "ease-of-use" could mean many things depending on the user interface and logical flow of functionality, etc…

Multinomial regression

- Introduce and illustrate multinomial regression
  - Model definition and interpretation
  - Model assessment
- Use Minitab to fit and assess the model
- Use model to identify important contributors



Visual Pain Scale

Ref: http://www.vcda.org/vi-visual-analog-pain-scale/

Various regression models

	Method	Input	Output	Other tools
Example 1	Linear regression	Single continuous	Single continuous	
Example 2	Multiple regression (including polynomial)	Multiple continuous and categorical	Single continuous	Regression tree; generalized additive model
Example 3	Logistic regression	Multiple continuous and categorical	Single binary	Classification tree; log linear model
Example 4	Multinomial regression	Multiple continuous and categorical	Single multinomial	Classification tree; log linear models
Server memory usage ~ f(# processed jobs)	Nonlinear regression	Multiple continuous and categorical	Single continuous	Regression tree; generalized additive model
Bluetooth signal ~ f(different wireless devices)	Nonparametric regression	Multiple continuous and categorical	Single continuous	
image features~ f(image markers, different background, resolution)	Multivariate regression	Multiple continuous	Multiple continuous	Partial least squares
duration to process order ~ f(order attributes, arrival time indicators)	Life time regression	Multiple continuous and categorical	Single continuous (life time)	
stock prices ~ f(company's financial indicators, competitor indicators, sector performance)	Robust regression	Multiple continuous and categorical	Single continuous	Quantile regression

What regression does not offer!!

- Correlation is not causation!!
- Bias in data selection (garbage in garbage out, missing information, only profile good customers, etc…)
  - Challenger o-ring failure
- Predict beyond scope of the data used to generate the regression model (exercise caution and common sense)
  - Fine print of any investment firm: "Past performance does not guarantee future returns"
- Regression model is merely a "close approximation" to the truth.
  - The best model still predicts with error.
  - Over-fit vs. under-fit
- Completely automated tool
  - There is no substitute for a well trained person to build and validate a model.

Remarks

- Regression is a powerful tool because finding and assessing a relationship is important in today's analytics. Uncovering insight in large amount of data in high dimensions is the norm rather than an exception.

- Regression offers insight into the contributing factors and the form of the relationship. Separating signal from noise and prioritizing contributors are 2 crucial benefits of regression.

- The type of regression depends on the problem. Complex problem requires complex solutions. Customized or targeted model offers focused solution.

- Regression has its limitations. Over the years, different solutions have been introduced to overcome these limitations. A skilled analyst is still the best solution today.

References

- Agresti, A. (2013), Categorical Data Analysis, 3rd edition, Wiley, NY.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, N.Y.
- Draper, N. and Smith, H. (1981), *Applied Regression Analysis*, Wiley, N.Y.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, N.Y.
- Louviere, J.J., Hensher, D.A. and Swait, J.D. (2000), *Stated Choice Methods: Analysis and Application*, Cambridge University Press, Cambridge, U.K.
- McCullaugh, P. and Nelder, J. (1989), *Generalized Linear Models*, Chapman and Hall, N.Y.
- Neter, J., Wasserman, W., and Kunter, M.H. (1989), *Applied Linear Regression Models*, Irwin, N.Y.
- Stokes, M.E., Davis, C.S., and Koch, G.G. (1995), *Categorical Data Analysis: using SAS system*, SAS Institute, Cary, NC.
- Weisberg, S. (1985), *Applied Linear Regression*, Wiley, N.Y.

Support Vector Machines

Support vector machines

Similar to linear regression they aim to find a hyperplane that linearly separates data points belong to different classes

In addition SVMs aim to find the hyperplane that is least likely to over fit the training data

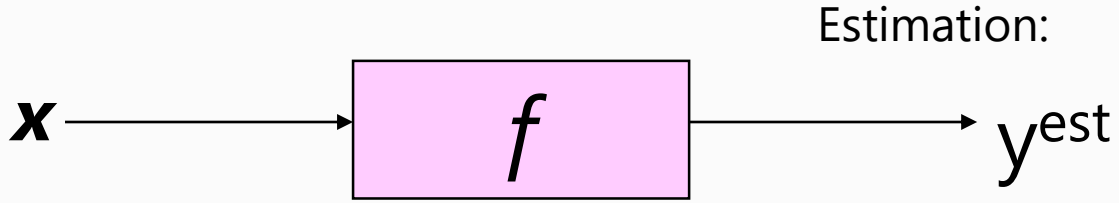
- By design, similar to pruning in Decision Trees, SVMs attempt to regulate the hypothesis space to ensure good performance on validation set…

- Fast in the nonlinear case
  - Use a mathematical trick to avoid creating "pseudo-attributes"
  - The nonlinear space is created implicitly

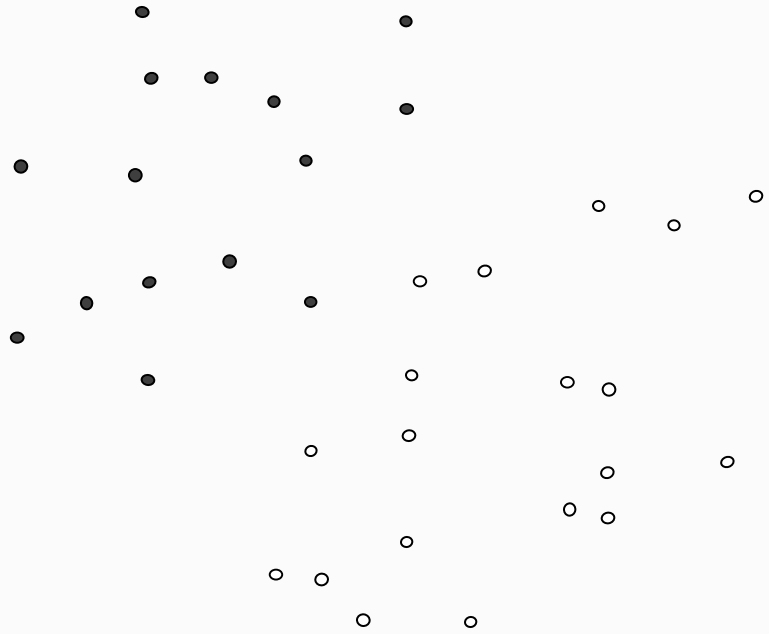
SVM, in a nutshell…

- SVM views the input data as two sets of vectors in an n-dimensional space. It constructs a separating hyperplane in that space, one which maximizes the margin between the two data sets.
- To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane.
- A good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes.
- The vectors (points) that constrain the width of the margin are the support vectors.

Linear Classifiers



- denotes +1
- denotes -1

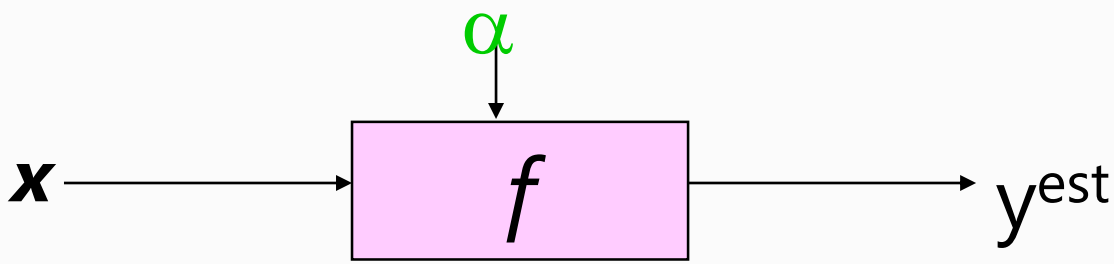


$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\ \mathbf{x} - b)$

w: weight vector
x: data vector

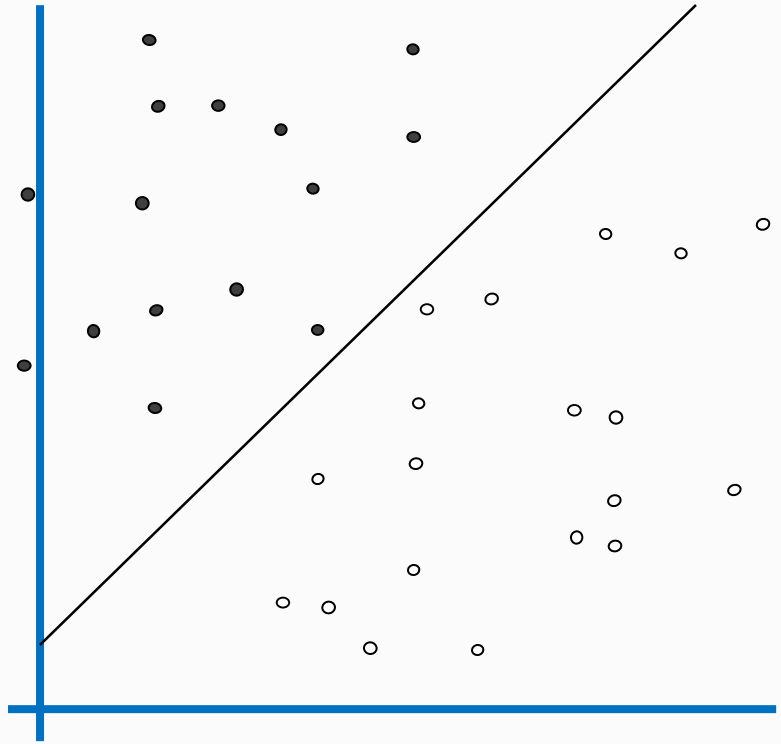
How would you classify this data?

Linear Classifiers



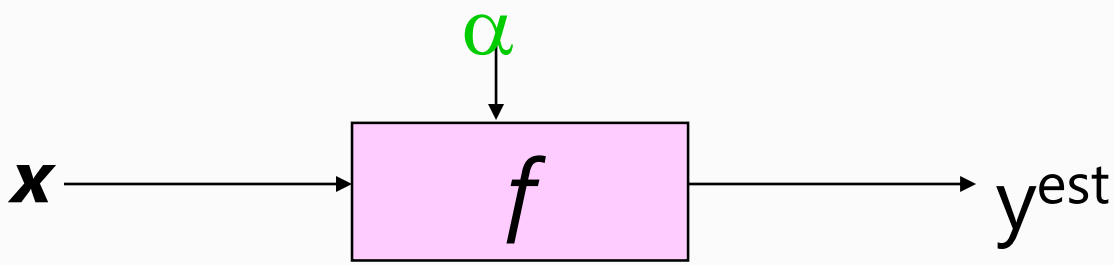
- denotes +1
- denotes -1

$$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\ \mathbf{x} - b)$$



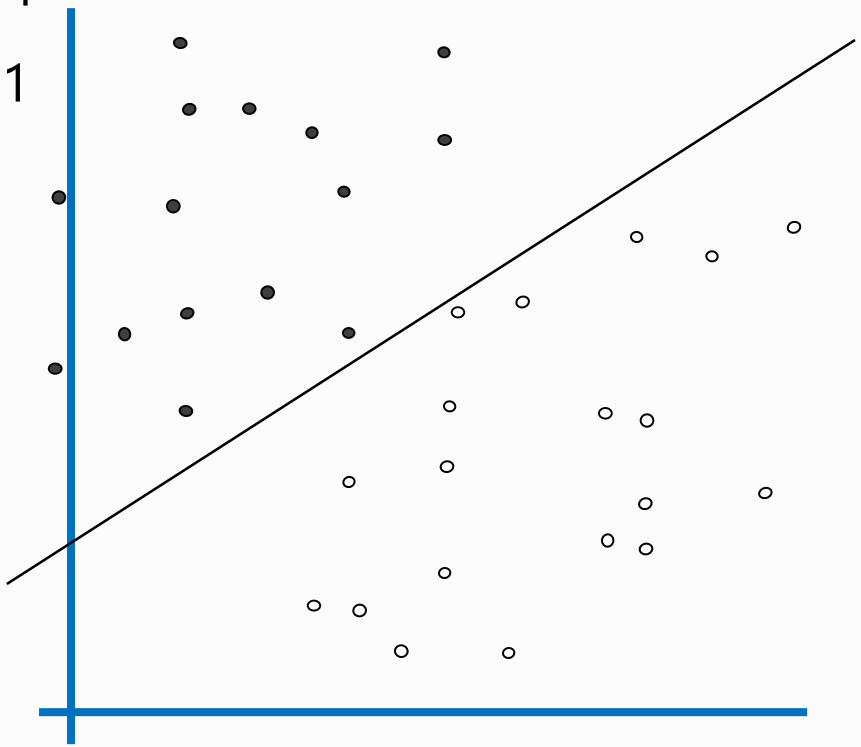
How would you classify this data?

Linear Classifiers



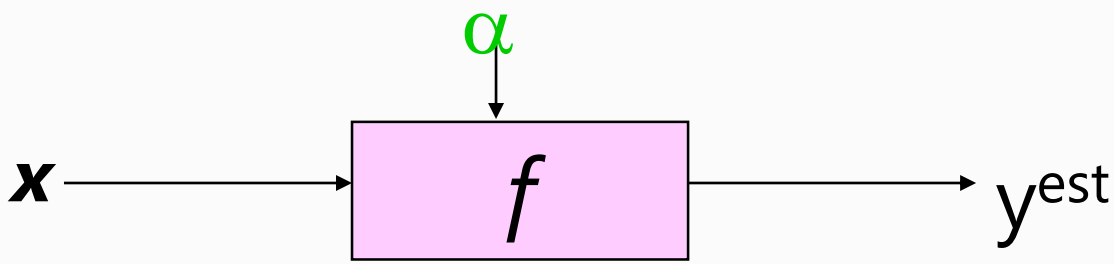
$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} - b)$

- denotes +1
- denotes -1

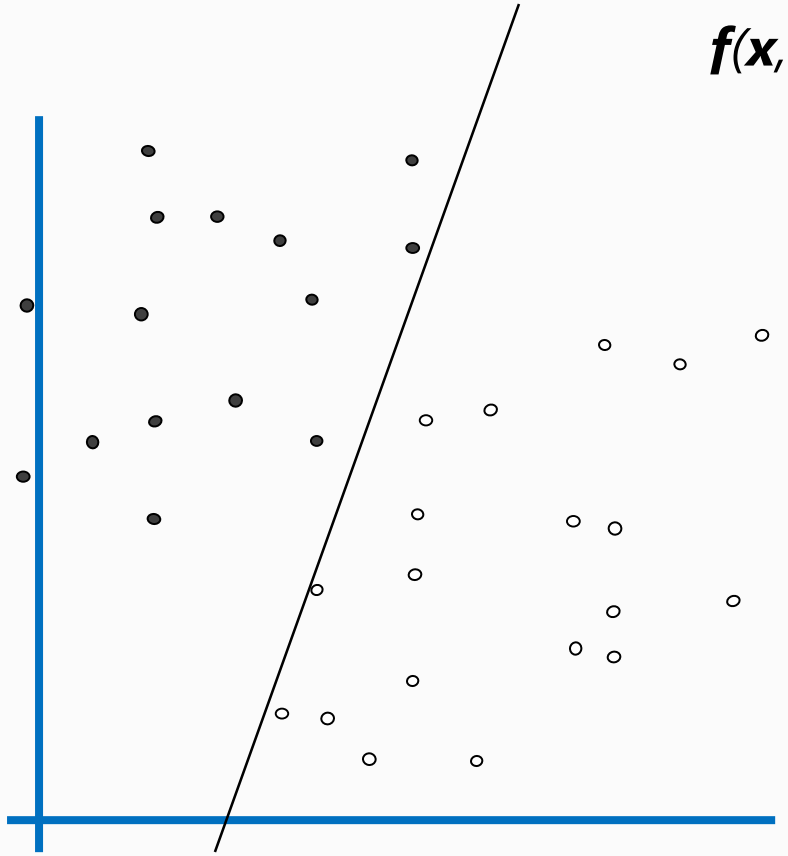


How would you classify this data?

Linear Classifiers



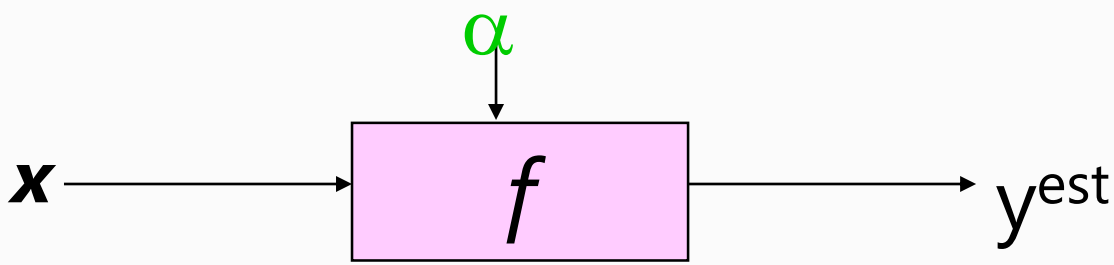
- denotes +1
- denotes -1



$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} - b)$

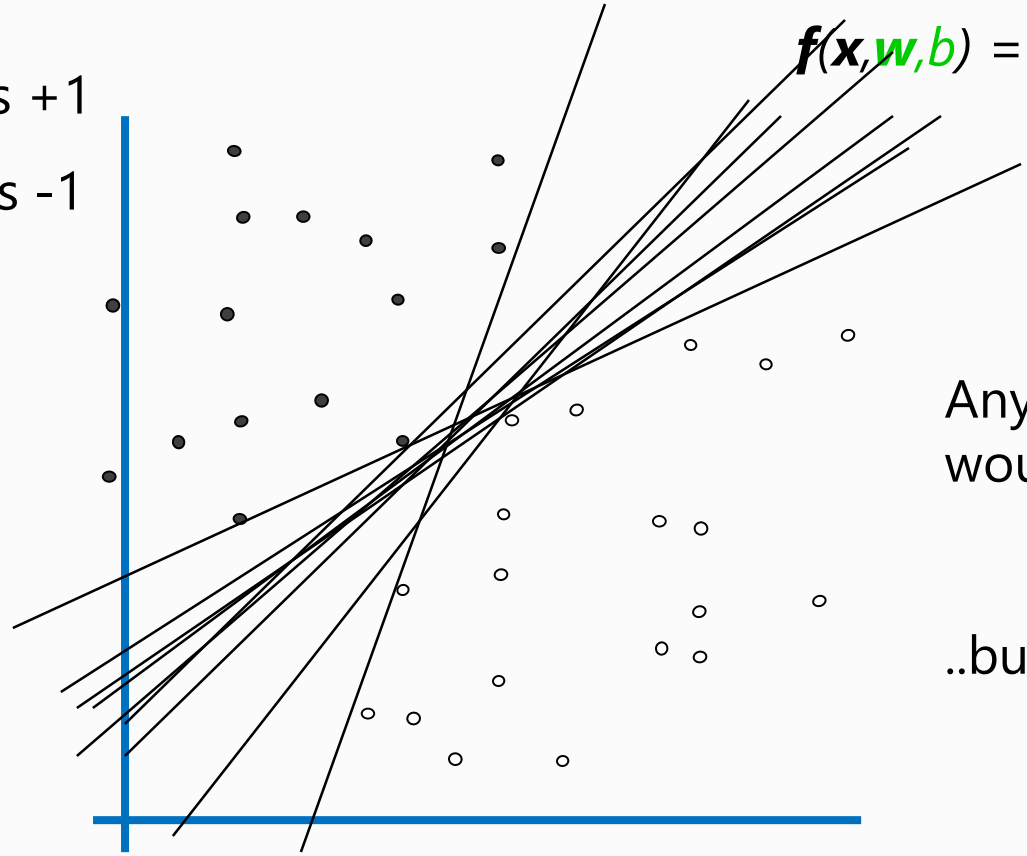
How would you classify this data?

Linear Classifiers



- denotes +1
- denotes -1

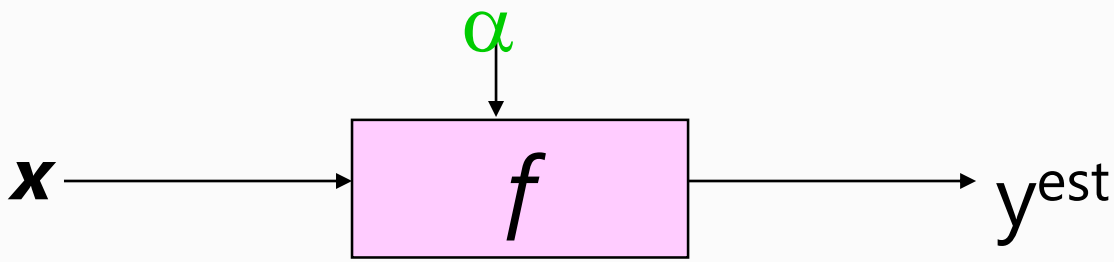
$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} - b)$



Any of these would be fine..

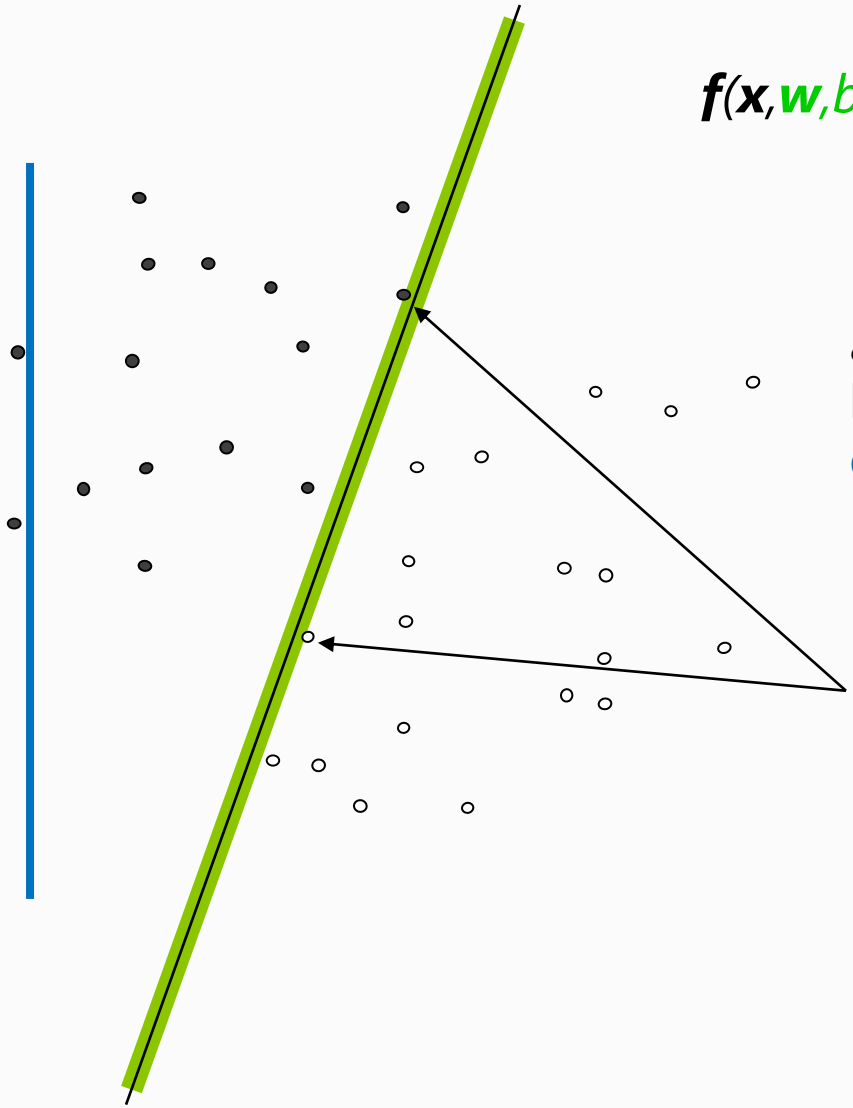
..but which is best?

Classifier Margin



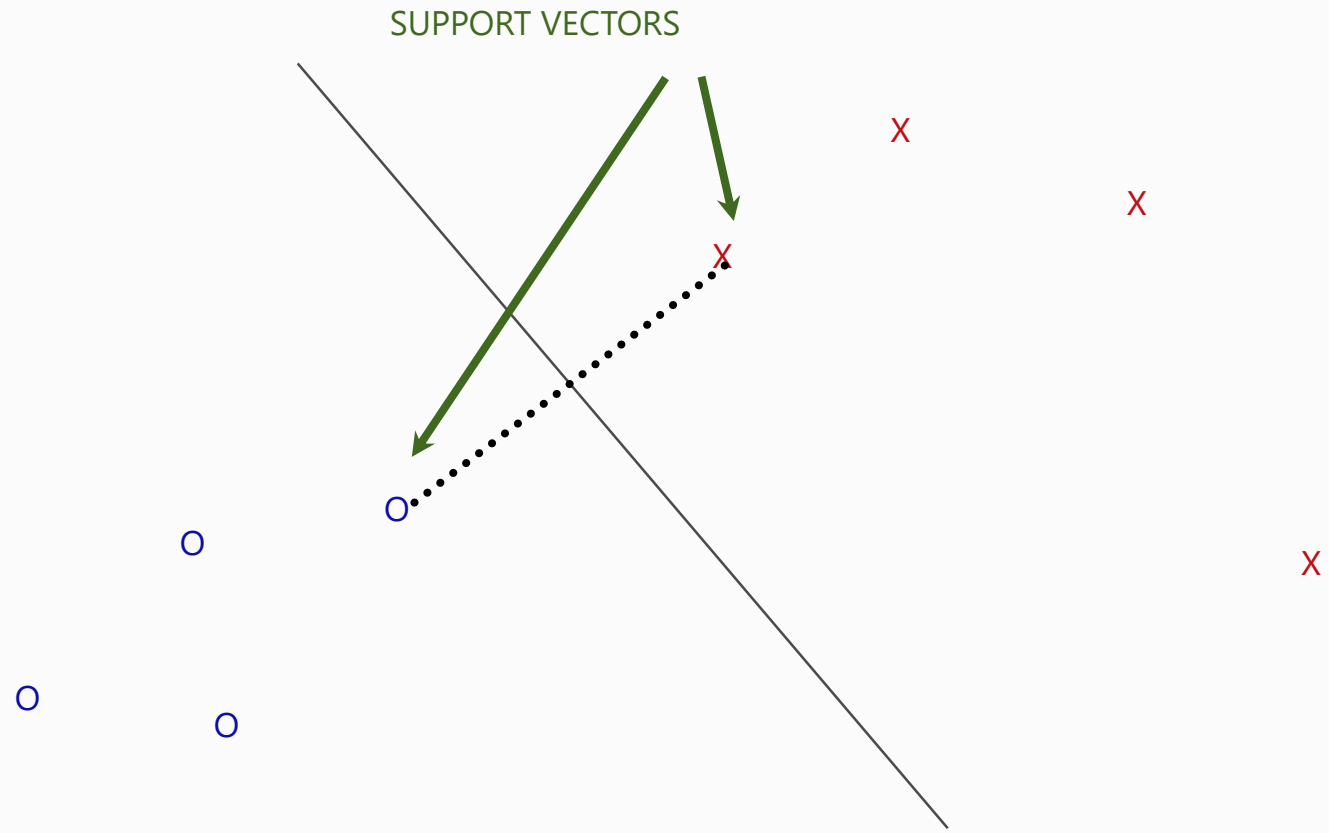
$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} - b)$

- denotes +1
- denotes -1

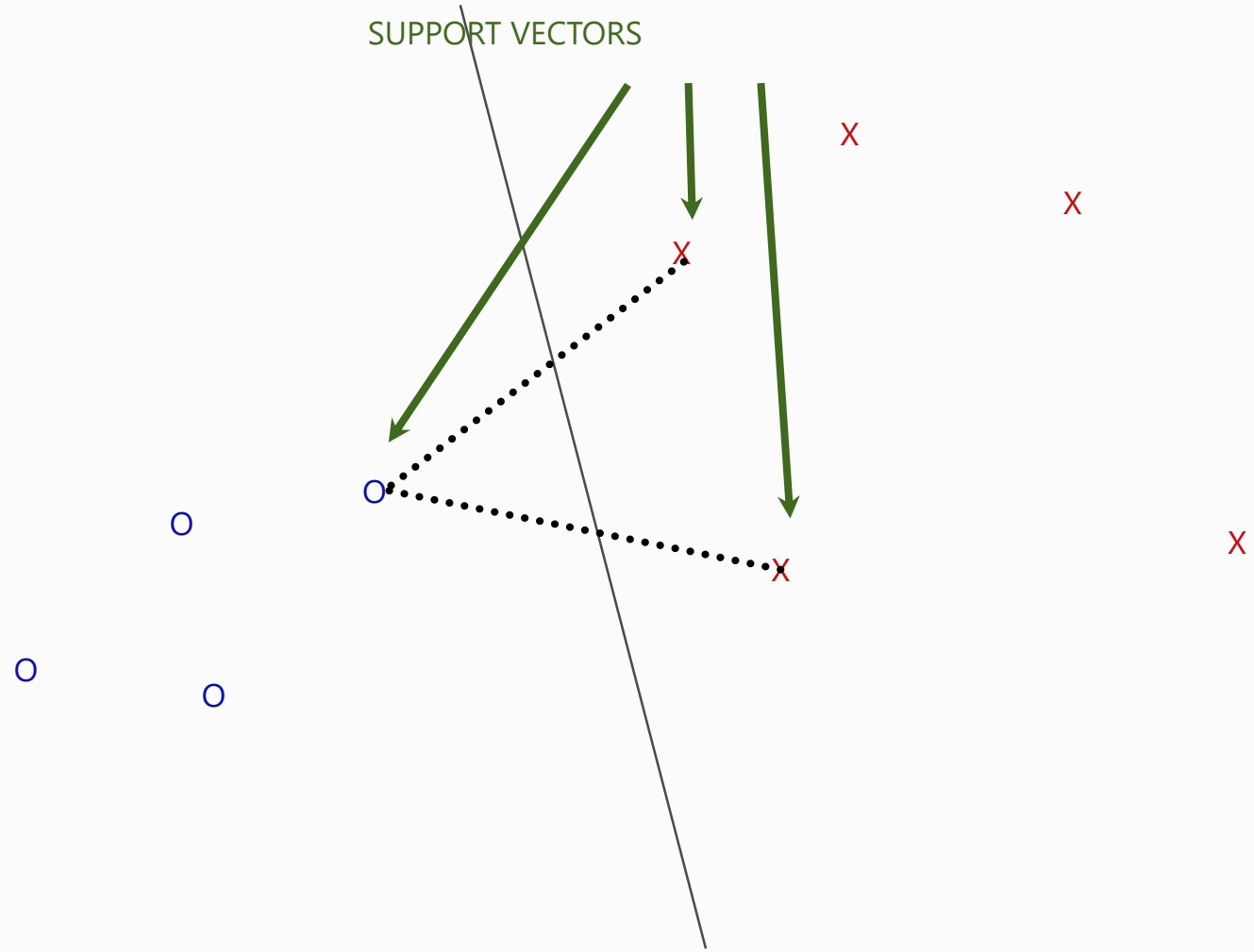


Define the margin of a linear classifier as the width that the boundary could be increased by before hitting a datapoint.

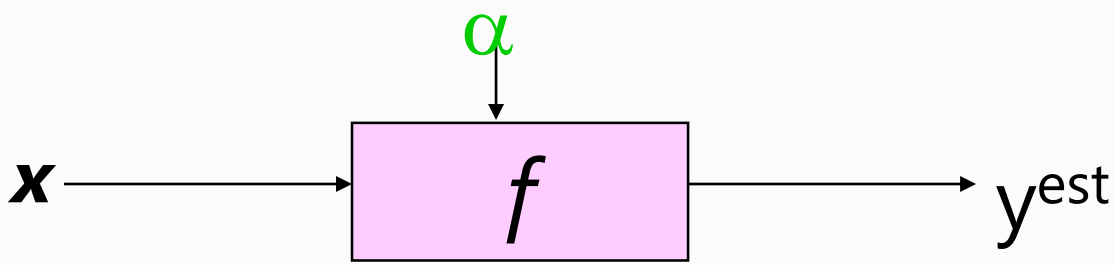
Geometric Intuition



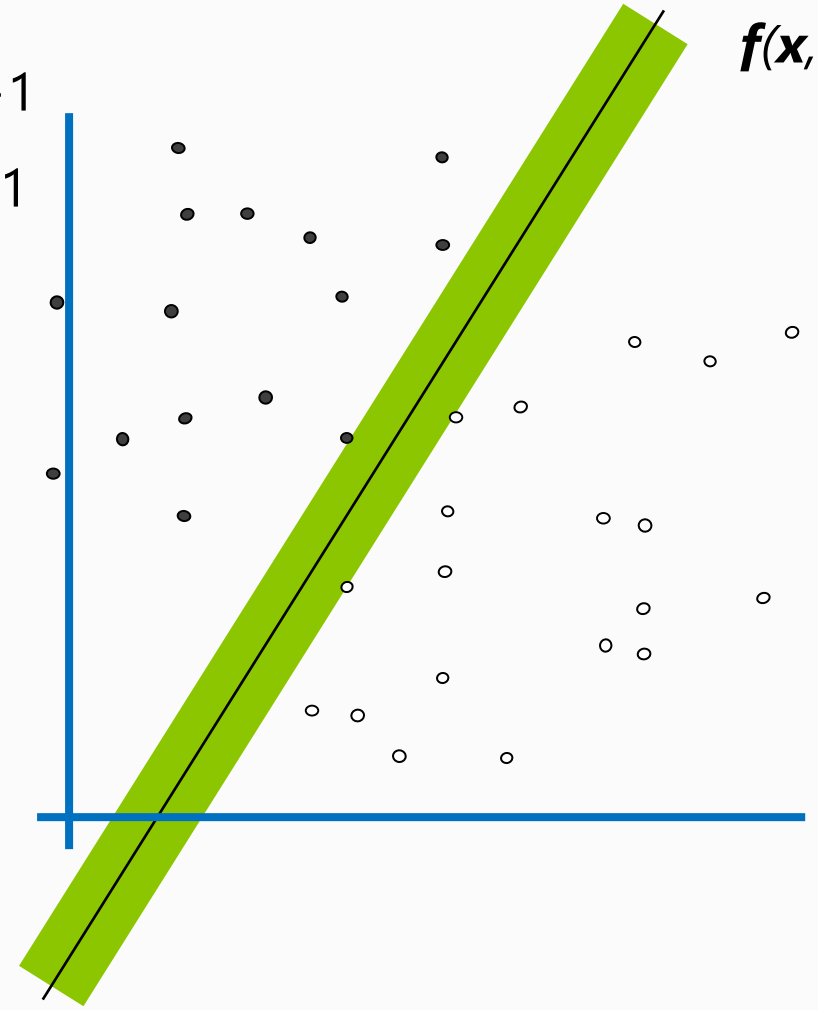
Geometric Intuition



Maximum Margin



- denotes +1
- denotes -1



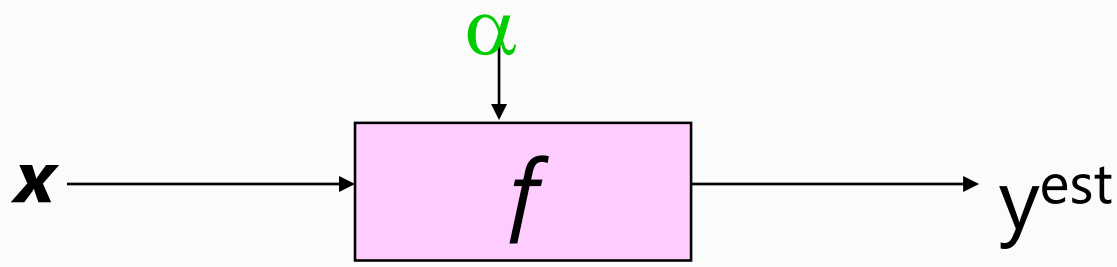
$f(x,w,b) = sign(w.\ x - b)$

The maximum margin linear classifier is the linear classifier with the maximum margin.

This is the simplest kind of SVM (Called an LSVM)

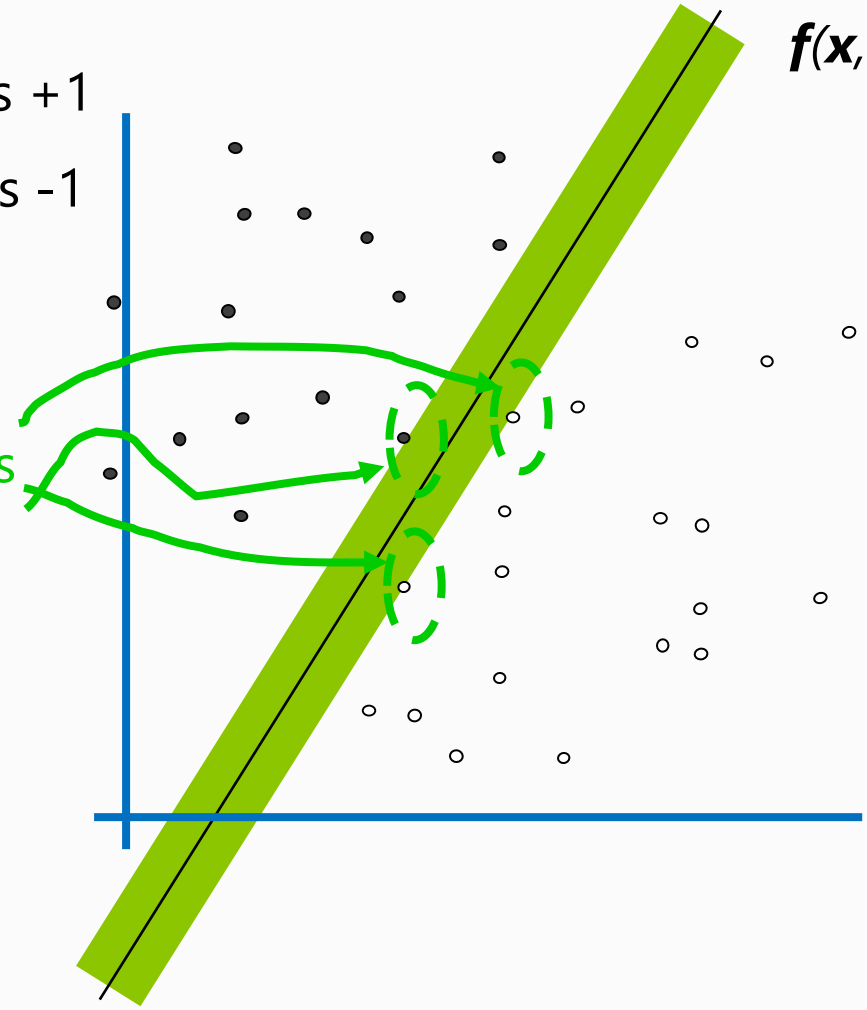
Linear SVM

Maximum Margin



$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}.\mathbf{x} + b)$

- denotes +1
- denotes -1



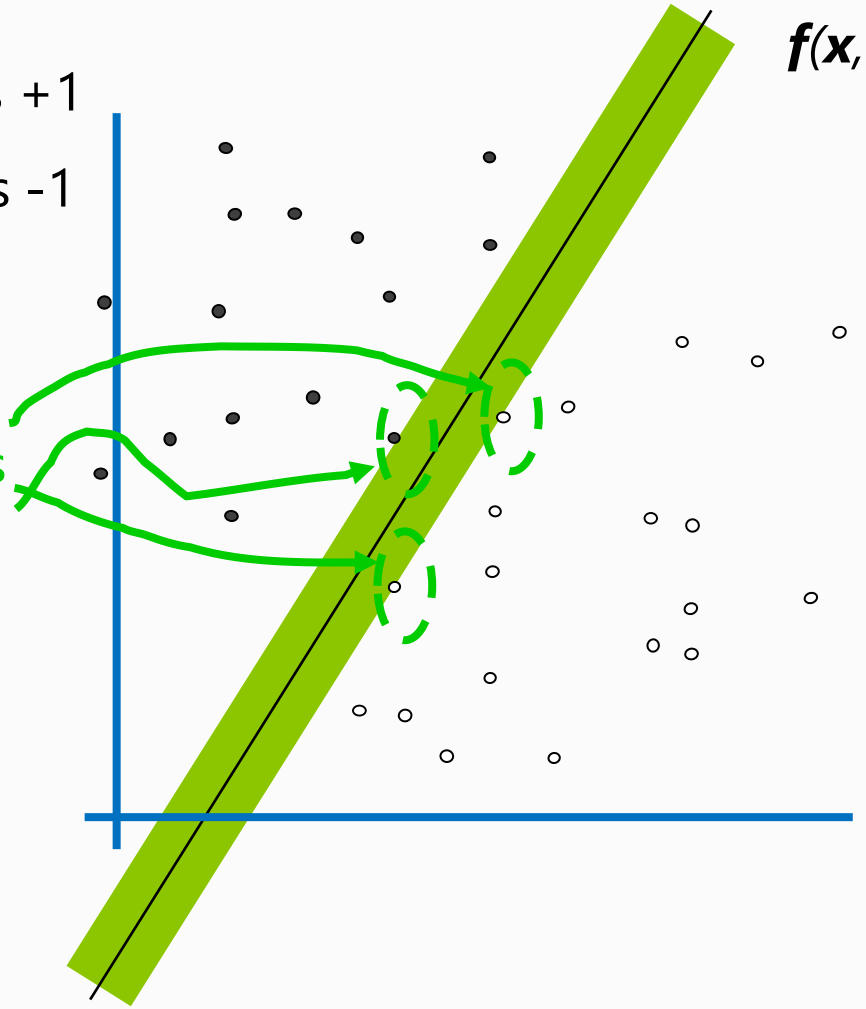
Support Vectors are those datapoints that the margin pushes up against

The maximum margin linear classifier is the linear classifier with the, um, maximum margin.

This is the simplest kind of SVM (Called an LSVM)

Linear SVM

Why Maximum Margin?



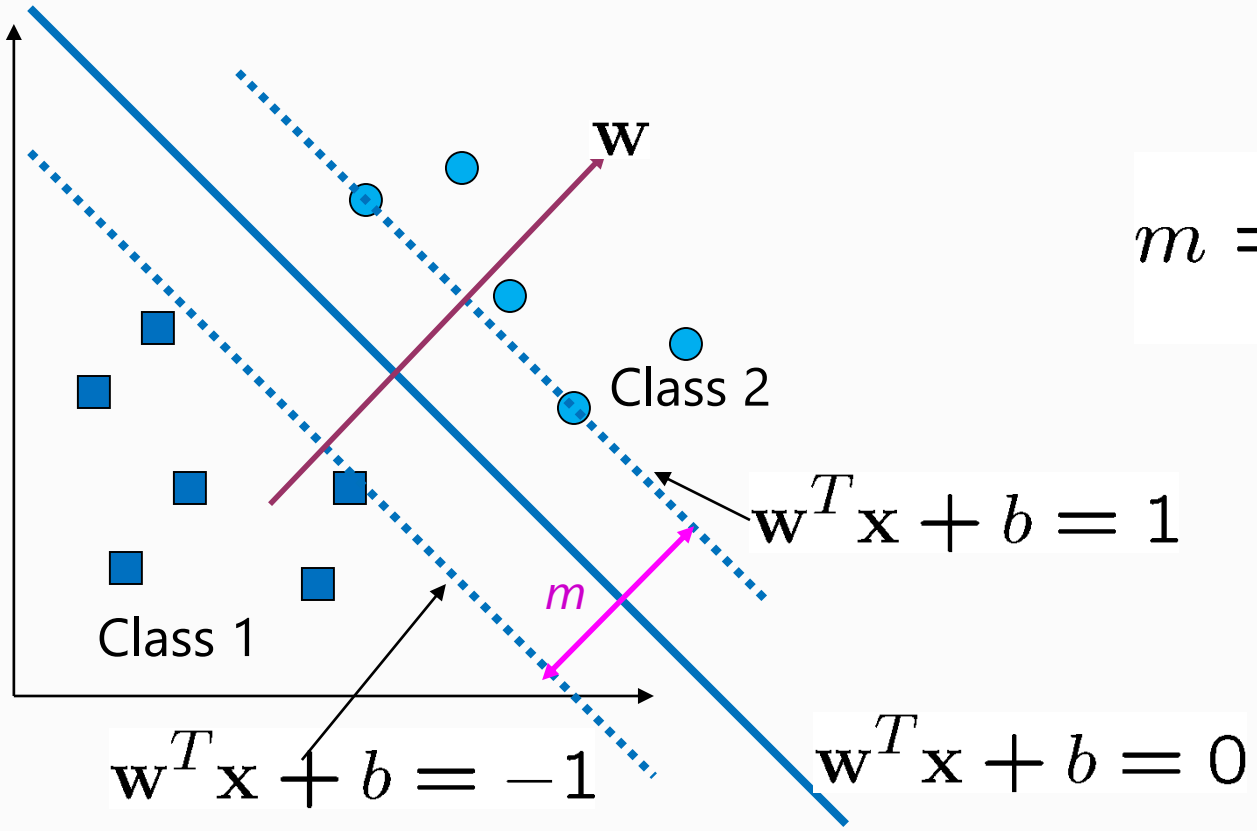
$$f(\boldsymbol{x},\boldsymbol{w},b) = sign(\boldsymbol{w}.\ \boldsymbol{x} - b)$$

1. Intuitively this feels safest.
2. If we've made a small error in the location of the boundary this gives us least chance of causing a misclassification.
3. LOOCV (leave one out cross validation) is easy since the model is immune to removal of any nonsupport-vector data points.
5. Empirically it works **very** well.

Large-Margin Decision Boundary

The decision boundary should be as far away from the data of both classes as possible

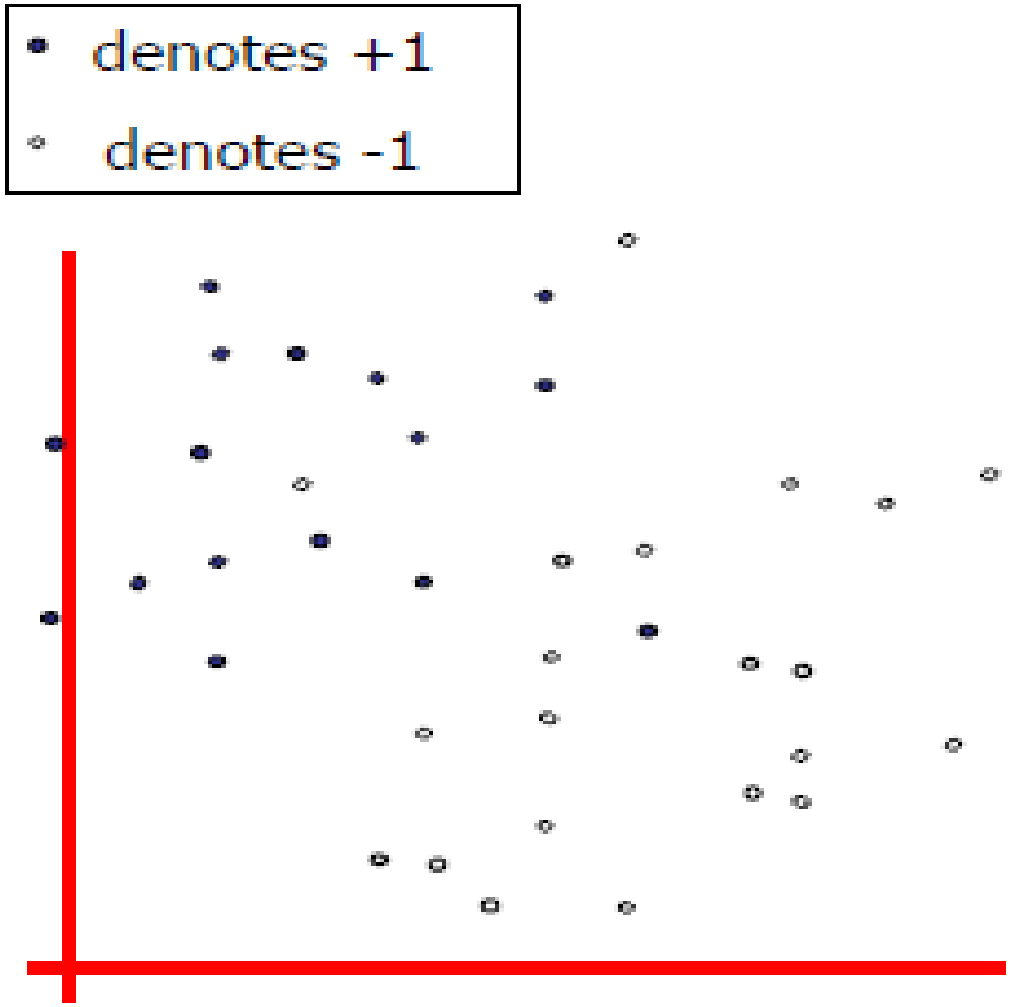
- We should maximize the margin, *m*
- Distance between the origin and the line $\mathbf{w}^t\mathbf{x}=-b$ is $b/||\mathbf{w}||$



$$m = \frac{2}{||\mathbf{w}||}$$

Uh-oh!

This is going to be a problem!



Slack Variables

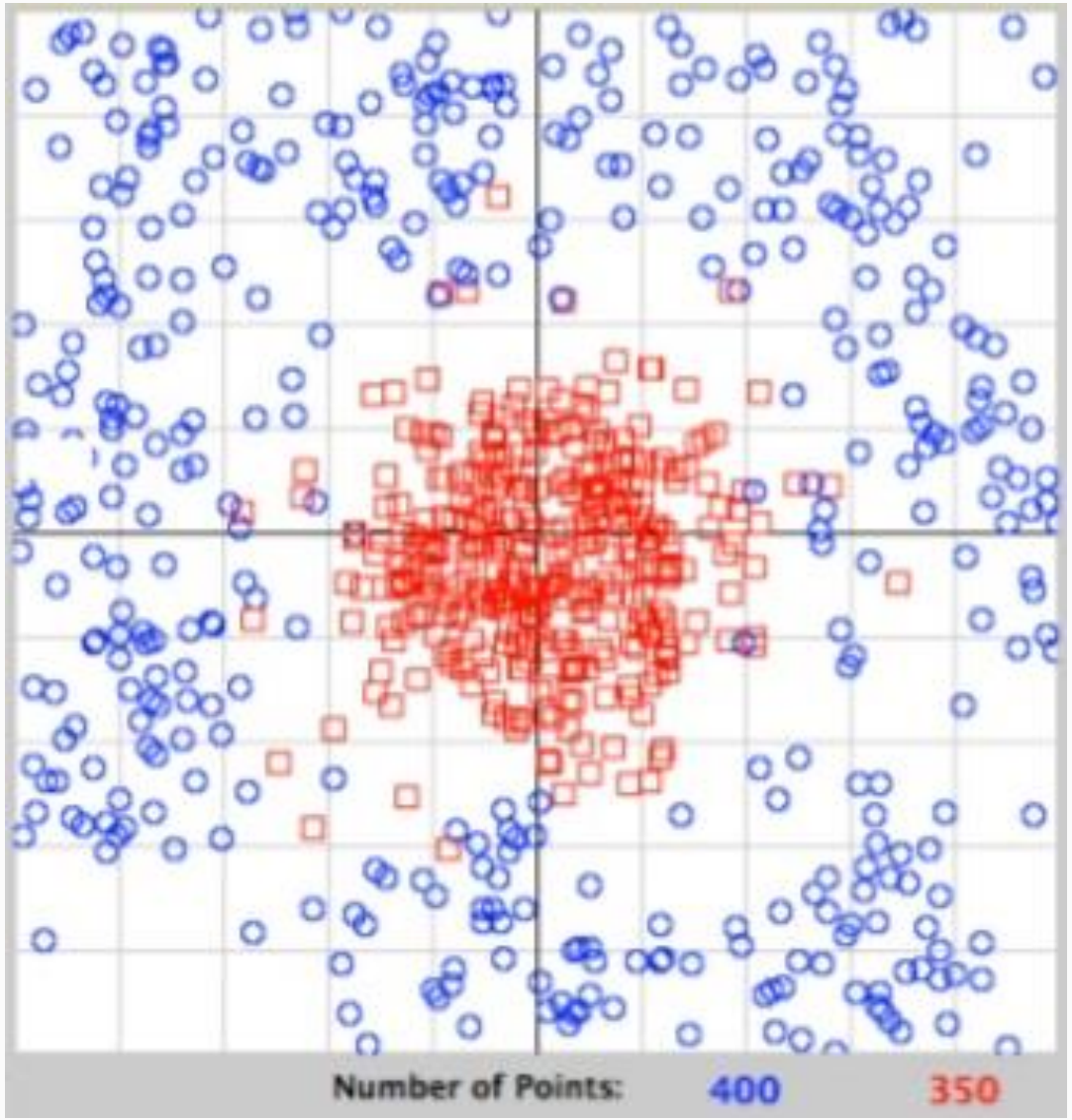
- SVMs are made robust by adding 'slack variables' that allow training error to be non-zero;
- One for each data point. Slack variable == 0 for correctly classified points

Noise

- Have assumed that the data is separable (in original or transformed space)
- Can apply SVMs to noisy data by introducing a "noise" parameter *C*
- C bounds the influence of any one training instance on the decision boundary
- Still a quadratic optimization problem
- Have to determine C by experimentation

The Kernel Trick

What if the data looks like that below?...

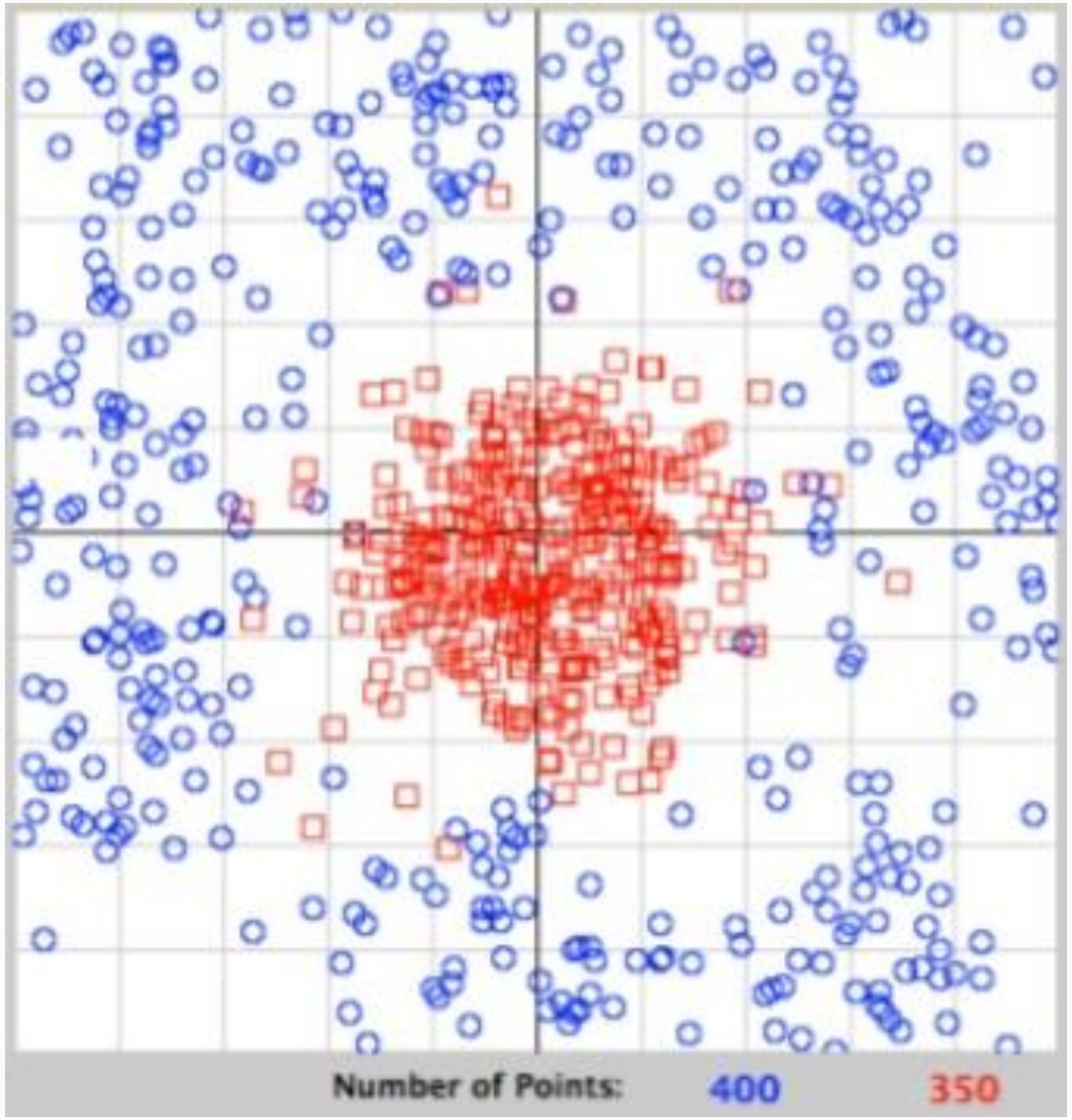


The simplest way to divide two groups is with a straight line, flat plane or an N-dimensional hyperplane. But what if the points are separated by a nonlinear region?

Rather than fitting nonlinear curves to the data, SVM handles this by using a *kernel function* to map the data into a different space where a hyperplane can be used to do the separation.

The Kernel Trick

What if the data looks like that below?...



The kernel trick allows you to use SVMs with non-linear separators

Different kernels

- Polynomial
- Gaussian
- Exponential
- …

The Kernel Trick

Everything you need to know in one slide…

Sometimes it improves your classifier if you map (i.e. transform) your data into a different feature space.

- Example 1: you have nonlinear data but want to use a linear classifier
- Example 2: your original data aren't numbers – for example they could be sequences

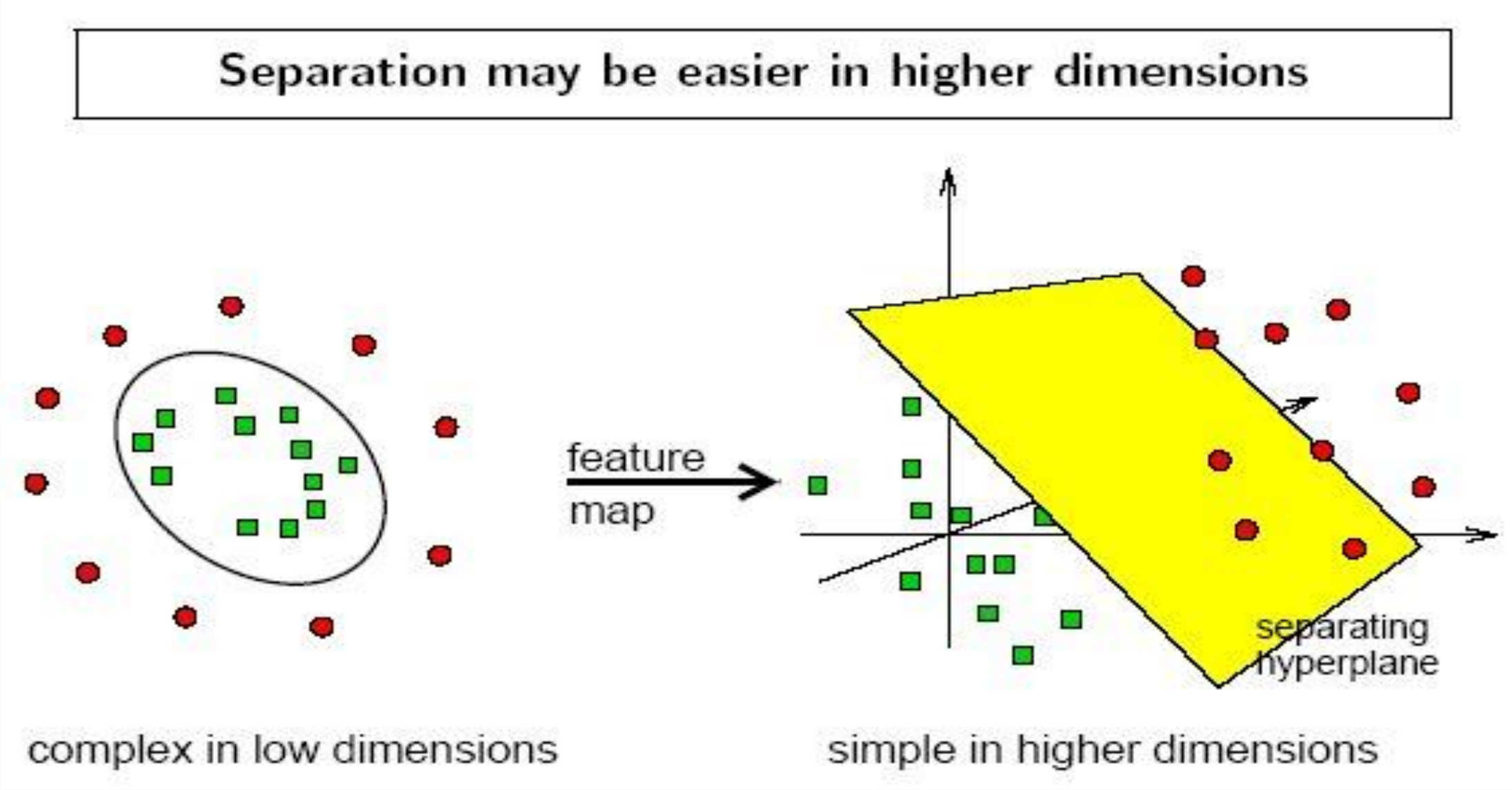
Kernel methods map your original data into a different space so that you can use linear classifiers.

This mapping can often substantially increases the number of features to consider.

- This can be problematic as your number of dimensions grows.

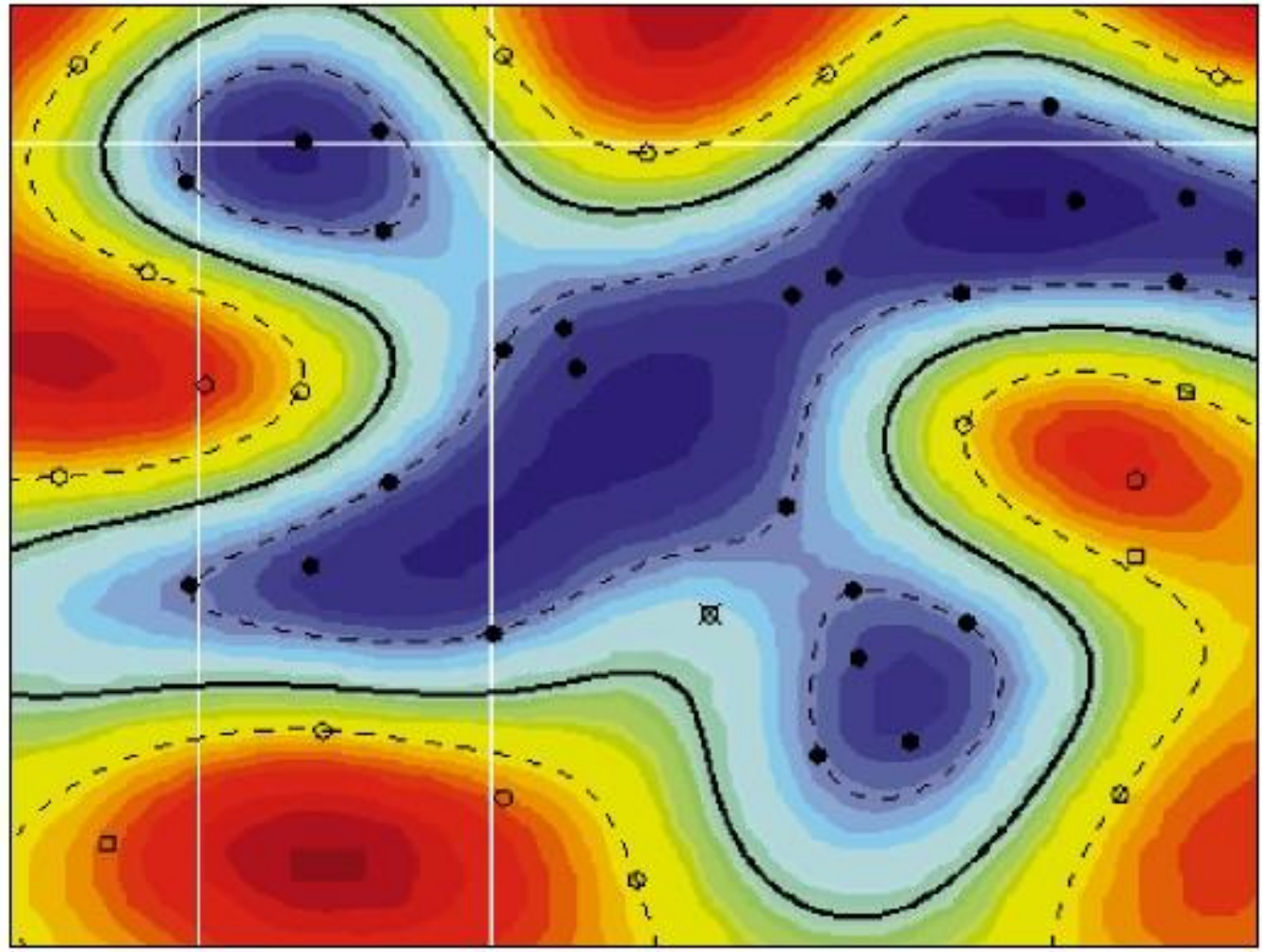
The "Kernel Trick" addresses this by putting a cap on the feature explosion so that the complexity of your classifier increases only linearly with the size of your original data.

Nonlinear Support Vector Machines



Nonlinear Support Vector Machines

The kernel function transforms the data into a higher dimensional space to make it possible to perform the separation.



Choosing the Kernel Function

Probably the most tricky part of using SVM

RBF is a good first option…

Depends on your data—try several.

- *Kernels have even been developed for nonnumeric data like sequences, structures, and trees/graphs.*

May help to use a combination of several kernels.

Don't touch your evaluation data while you're trying out different kernels and parameters.

– Use cross-validation for this if you're short on data

Type of Kernel	Inner product kernel $K(\vec{x}, \vec{x}_i),\ i = 1, 2, \ldots, N$	Comments
Polynomial Kernel	$K(\vec{x}, \vec{x}_i) = \left(\vec{x}^T \vec{x}_i + \theta\right)^d$	Power p and threshold θ is specified a priori by the user
Gaussian Kernel	$K(\vec{x}, \vec{x}_i) = e^{-\frac{1}{2\sigma^2}\|\vec{x}-\vec{x}_i\|^2}$	Width σ^2 is specified a priori by the user
Sigmoid Kernel	$K(\vec{x}, \vec{x}_i) = tanh(\eta\, \vec{x}\, \vec{x}_i + \theta)$	Mercer's Theorem is satisfied only for some values of η and θ
Kernels for Sets	$K(\chi, \chi') = \sum_{i=1}^{N_\chi} \sum_{j=1}^{N_{\chi'}} k(x_i, x'_j)$	Where $k(x_i, x'_j)$ is a kernel on elements in the sets χ, χ'
Spectrum Kernel for strings	count number of substrings in common	It is a kernel, since it is a dot product between vectors of indicators of all the substrings.

Table 1: Summary of Inner-Product Kernels [Hay98]

Complexity of the optimization problem remains only dependent on the dimensionality of the input space and not of the feature space!

Sparse Data

- SVM algorithms speed up dramatically if the data is sparse (i.e. many values are 0)
- Why? Because they compute lots and lots of dot products
- Sparse data compute dot products very efficiently
- Iterate only over nonzero values
- SVMs can process sparse datasets with 10,000s of attributes

Doing multi-class classification

- SVMs can only handle two-class outputs (i.e. a categorical output variable with arity 2).
- What can be done?
- Answer: with output arity N, learn N SVM's
  - SVM 1 learns "Output==1" vs "Output != 1"
  - SVM 2 learns "Output==2" vs "Output != 2"

    ….
  - SVM N learns "Output==N" vs "Output != N"
- Then to predict the output for a new input, just predict with each SVM and find out which one puts the prediction the furthest into the positive region.

What you need to know…

- First, try logistic regression. Easy, fast, stable. No 'tuning' parameters.
- Equivalently, you can first try linear SVMs, but you need to tune 'C'
- If results are "good enough", stop
- Else, try SVMs with Gaussian kernels (RBF)

  Need to tune bandwidth, C – by using validation data…

Summary: Steps for Classification

- SVMs require vector of real numbers
  - Categorical variables → numeric data {R,G,B} → {0,0,1},…{1,0,0}
  - Scaling to the range [-1, +1] or [0,1]
- Select the kernel function to use
  - RBF is a reasonable first choice, two parameters C and γ
  - Grid search to identify best values for parameters
  - Use v-fold cross validation to ensure good performance on test data
- Unseen data can be classified using support vectors

Conclusion

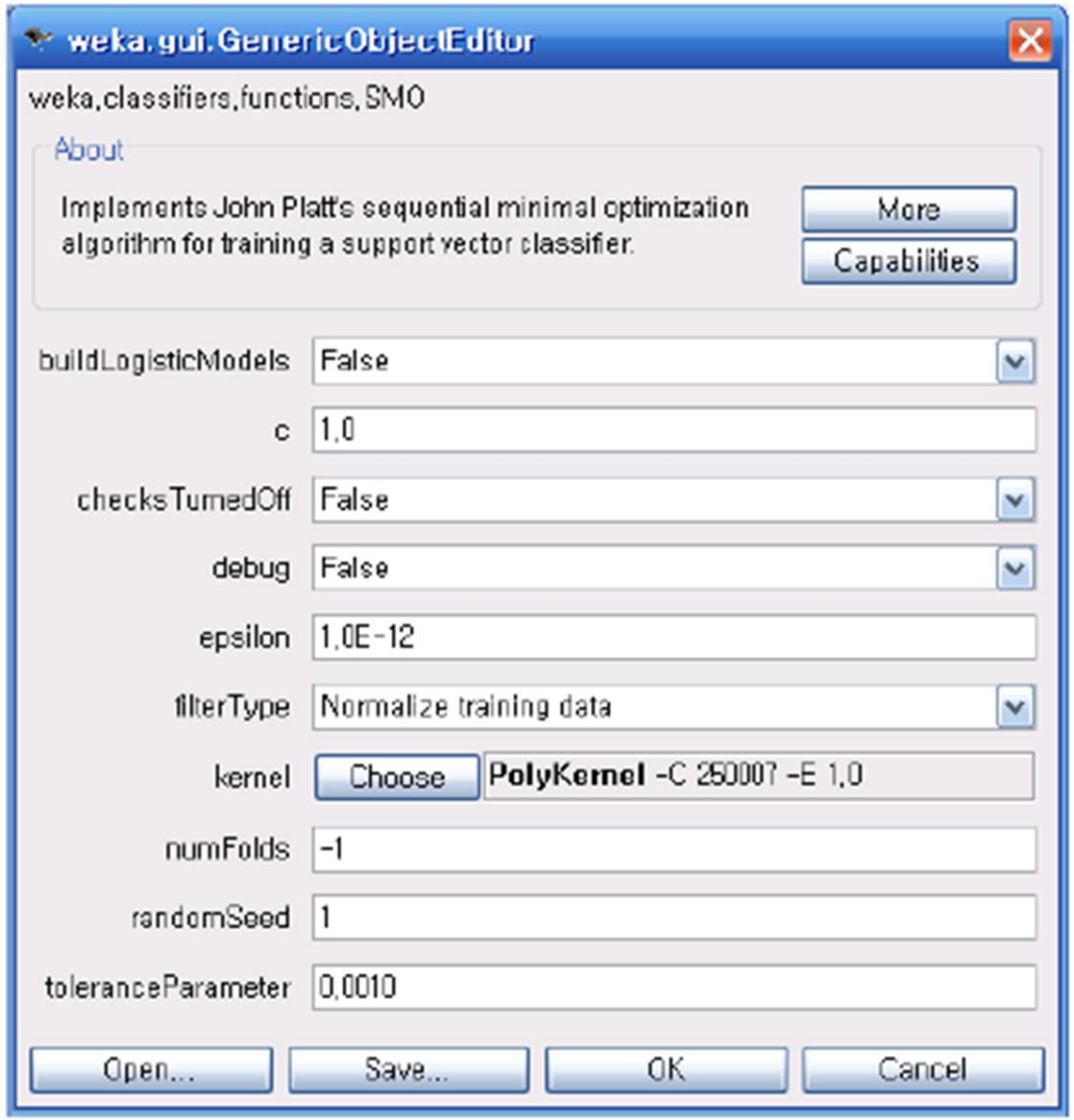
- SVMs balance between correctness and generalization
  - Decision boundaries
  - Margins
  - Support vector
- Two key concepts of SVM: maximize the margin and the kernel trick
- Many SVM implementations are available on the web for you to try on your data set!

Software

- A list of SVM implementation can be found at http://www.kernel-machines.org/software.html
- Some implementation (such as LIBSVM) can handle multi-class classification
- SVMLight is among one of the earliest implementation of SVM
- Several Matlab toolboxes for SVM are also available

In WEKA …

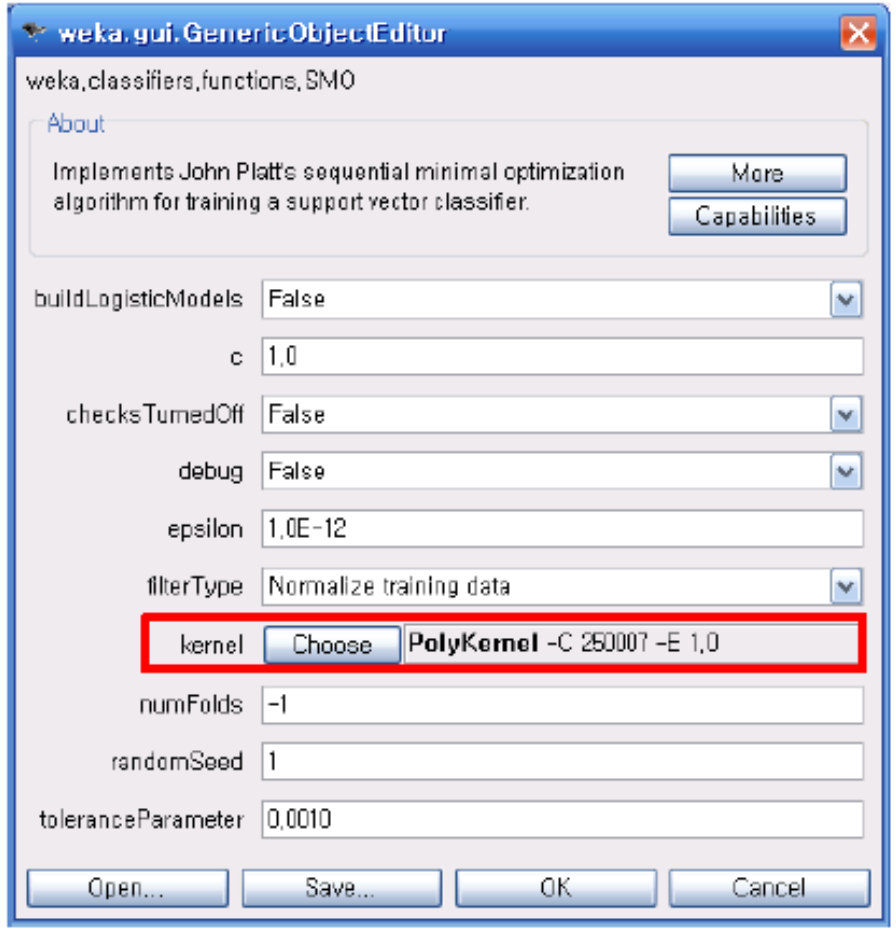
- Sequential minimal optimization (SMO) algorithm support vector *classification*
  - *weka.classifiers.functions.SMO*
- A Library for Support Vector Machines (libSVM)
  - *weka.classifiers.functions.libSVM*
- Sequential minimal optimization (SMO) algorithm support vector *regression*
  - *weka.classifiers.functions.SMOreg*



- buildLogisticModels: Whether to fit logistic models to the outputs (for proper probability estimates)
- numFolds: The number of folds for cross-validation used to generate training data for logistic models
- randomSeed: Random number seed for the cross-validation

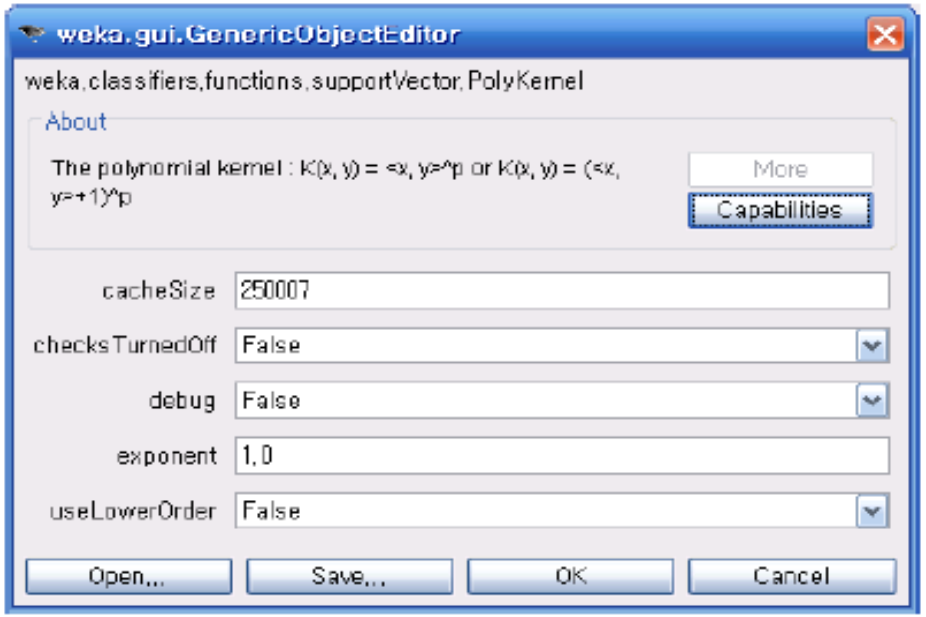
- c -- The complexity parameter C. It is the upper bound of alpha's
- filterType -- Determines how/if the data will be transformed.
- kernel -- The kernel to use.

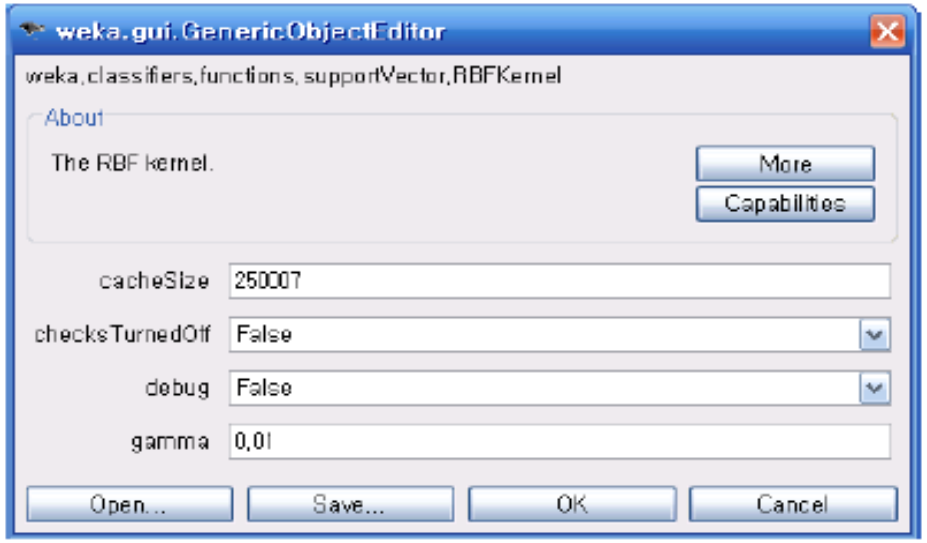
Parameter Setting Guide



- **Suggested**
  - buildLogisticModels: True
  - numFolds: 3 or 5
  - randomSeed: any value
  - C (complexity parameter, upper bound of alpha)
- **Your own choice**
  - filterType
  - Kernel and its subsequent parameters
  - Debug – if on, you can see intermediate results
- **Do not change!**
  - epsilon
  - checksTurnedOff
  - toleranceParameter

Parameter Setting Guide – Kernel

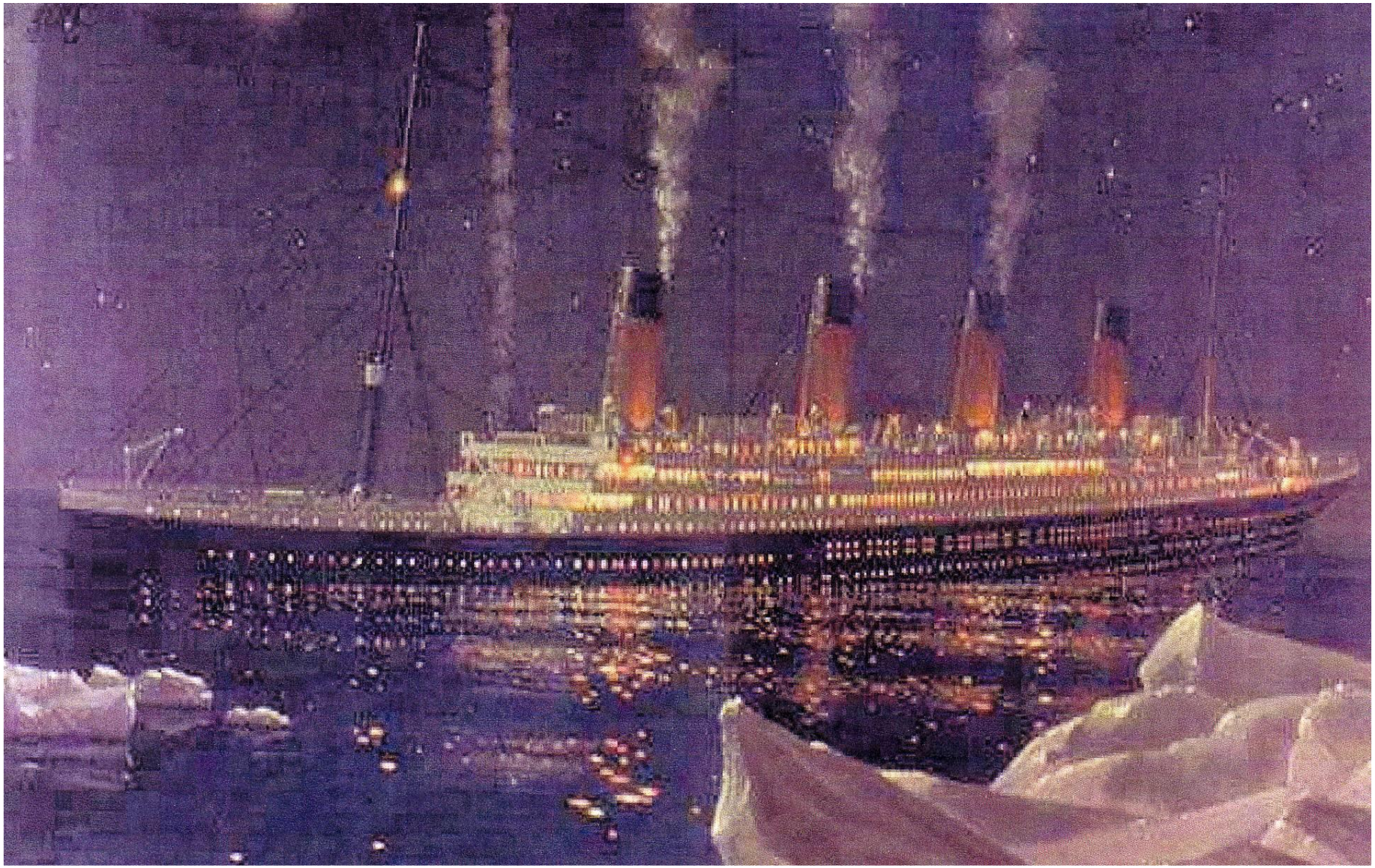




- **PolyKernel**
  - Try various exponents
    - Floating points are allowed
  - 1.0: linear kernel

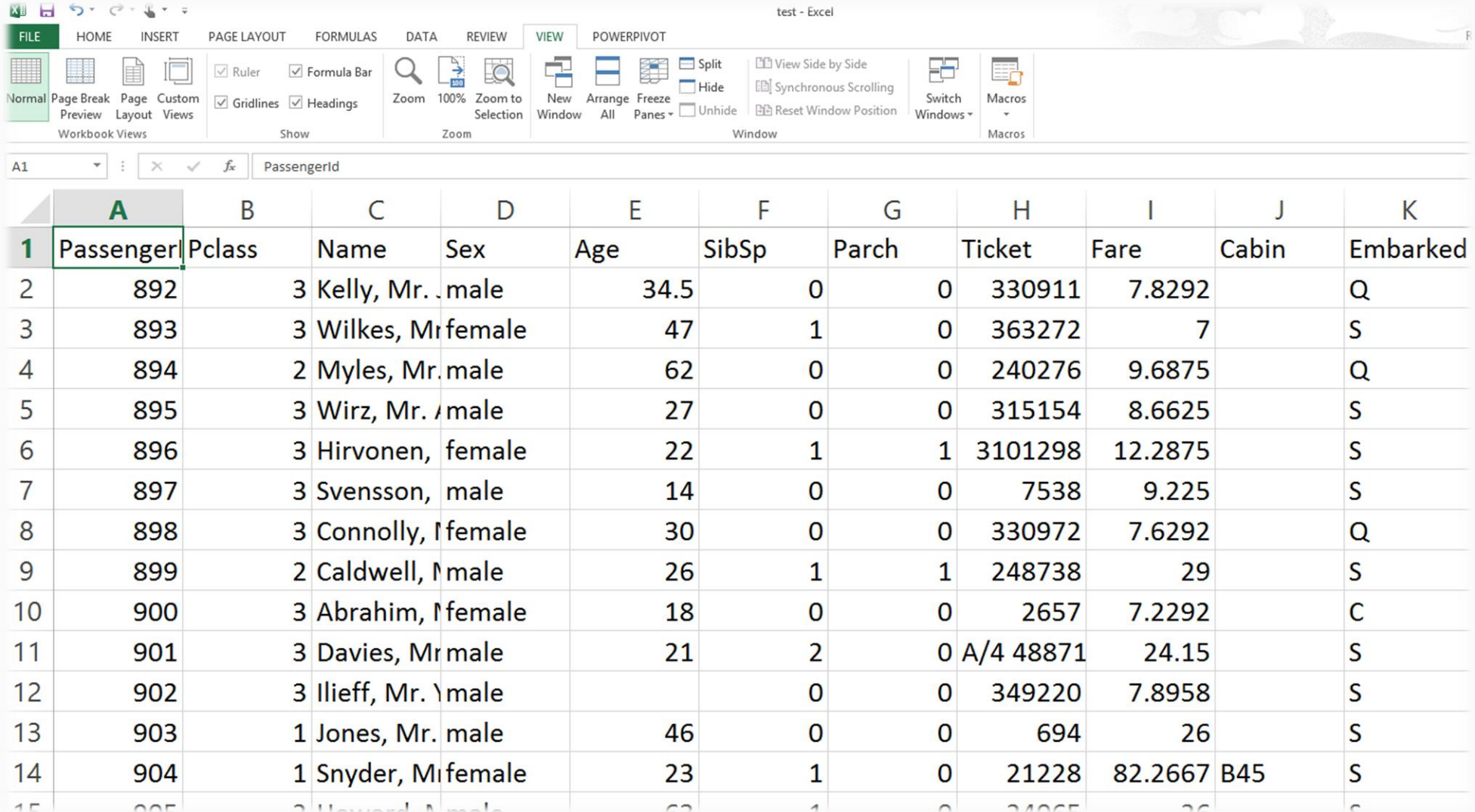
- **RBFKernel**
  - Try various gammas
  - Gamma value corresponds to the inverse of the variance (width of the kernel)

The Tragedy of the Titanic



First Steps…

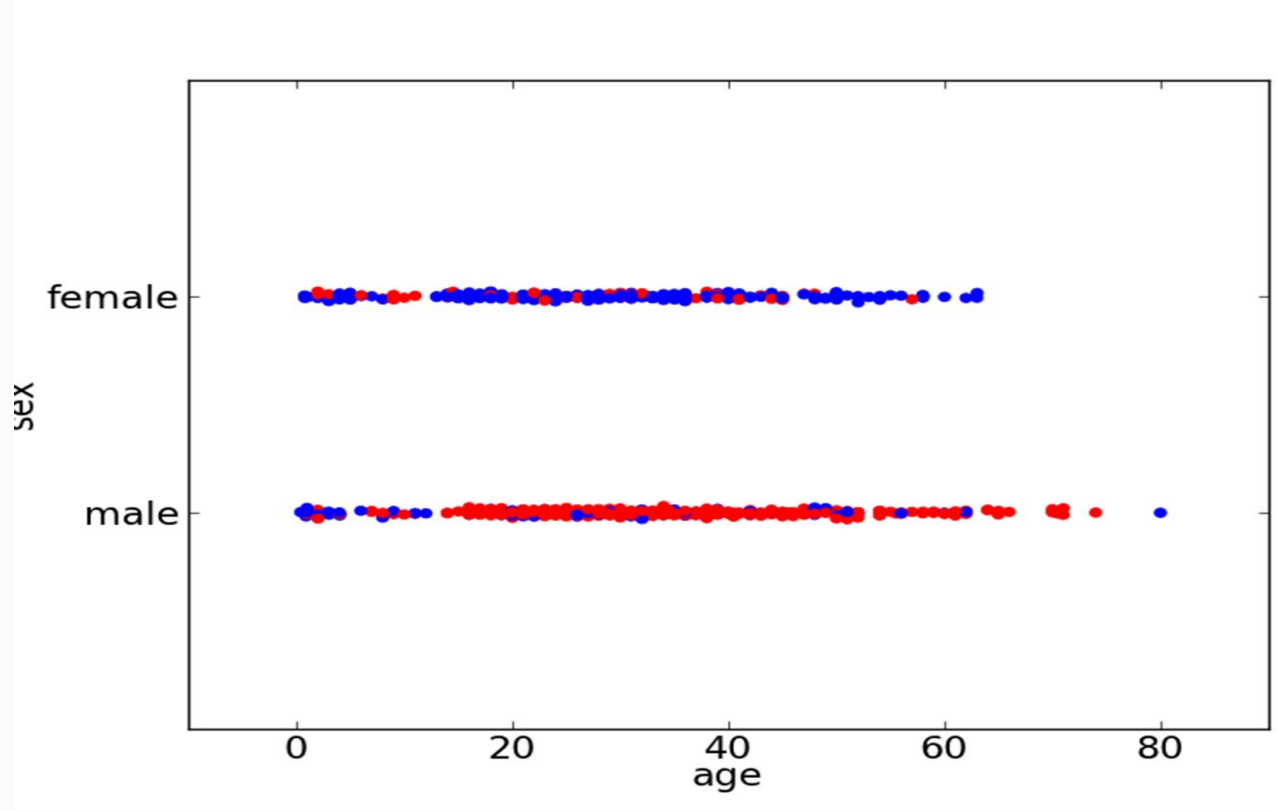
- Look at the data, exploratory data analysis: Excel, WEKA,…;



	A	B	C	D	E	F	G	H	I	J	K
1	PassengerI	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	892	3	Kelly, Mr. .	male	34.5	0	0	330911	7.8292		Q
3	893	3	Wilkes, Mr	female	47	1	0	363272	7		S
4	894	2	Myles, Mr.	male	62	0	0	240276	9.6875		Q
5	895	3	Wirz, Mr. /	male	27	0	0	315154	8.6625		S
6	896	3	Hirvonen,	female	22	1	1	3101298	12.2875		S
7	897	3	Svensson,	male	14	0	0	7538	9.225		S
8	898	3	Connolly, I	female	30	0	0	330972	7.6292		Q
9	899	2	Caldwell, I	male	26	1	1	248738	29		S
10	900	3	Abrahim, I	female	18	0	0	2657	7.2292		C
11	901	3	Davies, Mr	male	21	2	0	A/4 48871	24.15		S
12	902	3	Ilieff, Mr. \	male		0	0	349220	7.8958		S
13	903	1	Jones, Mr.	male	46	0	0	694	26		S
14	904	1	Snyder, Mi	female	23	1	0	21228	82.2667	B45	S

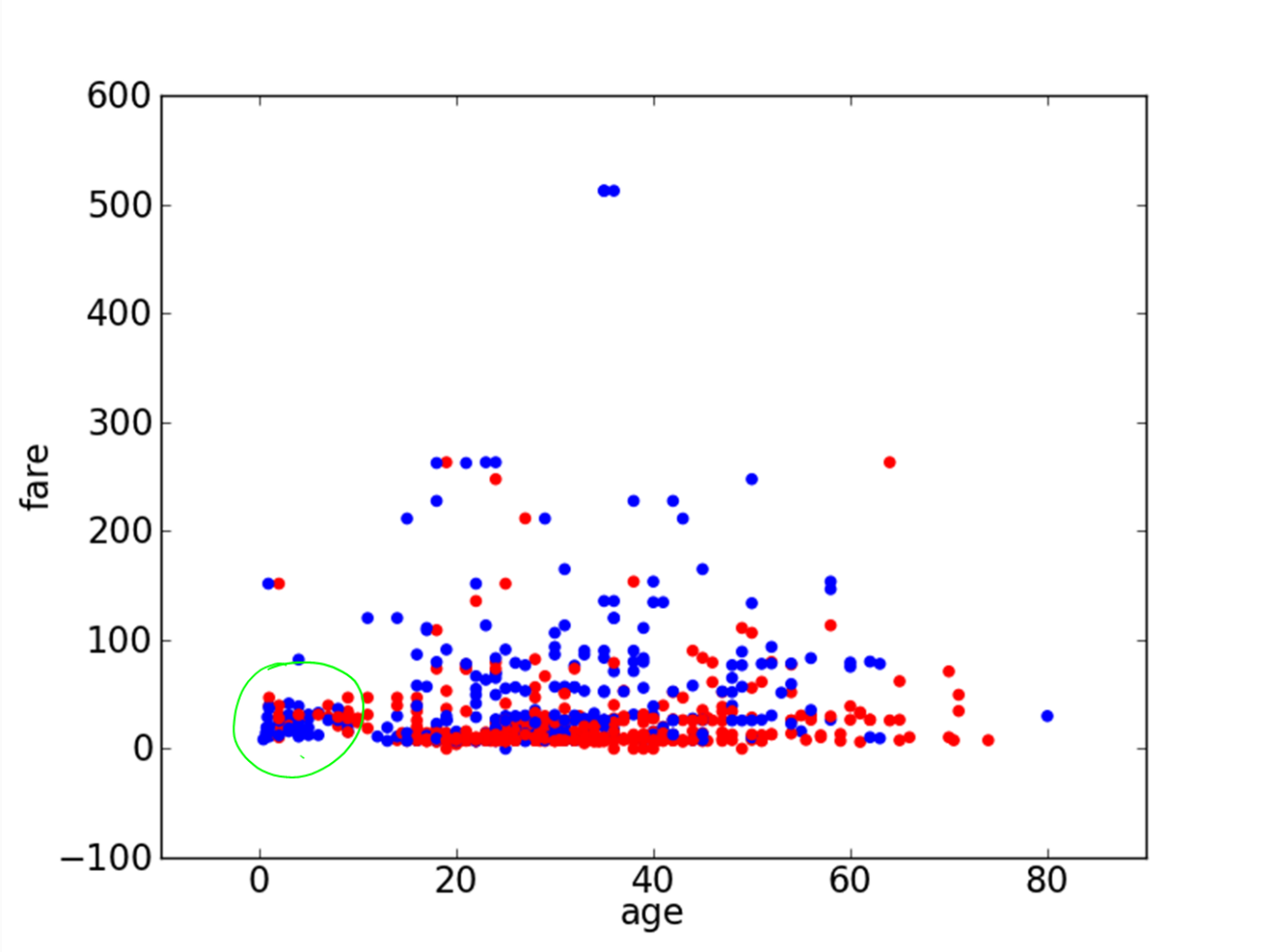
First Steps…

- Look at the data;
- Follow your intuition, *"women and children first…"*
  - *What percentage of Females survived, percentage of males survived?…*



First Steps…

- Other features good predictors?



A very naïve classifier

pclass	sex	age	sibsp	parch	fare	cabin	embarked
1	female	35	1	0	53.1	C123	S

Does the new data point x^* ***exactly*** match a previous point x_i?

If so, assign it to the same class as x_i

Otherwise, just guess.

This is the "rote" classifier

A minor improvement

pclass	sex	age	sibsp	parch	fare	cabin	embarked
1	female	35	1	0	53.1	C123	S

Does the new data point x^* match a set of previous points x_i on **some specific attribute**?

If so, take a vote to determine class.

Example: If most females survived, then assume every female survives

But there are lots of possible rules like this.
And an attribute can have more than two values.

If most people under 4 years old survive, then assume everyone under 4 survives
If most people with 1 sibling survive, then assume everyone with 1 sibling survives

How do we choose?

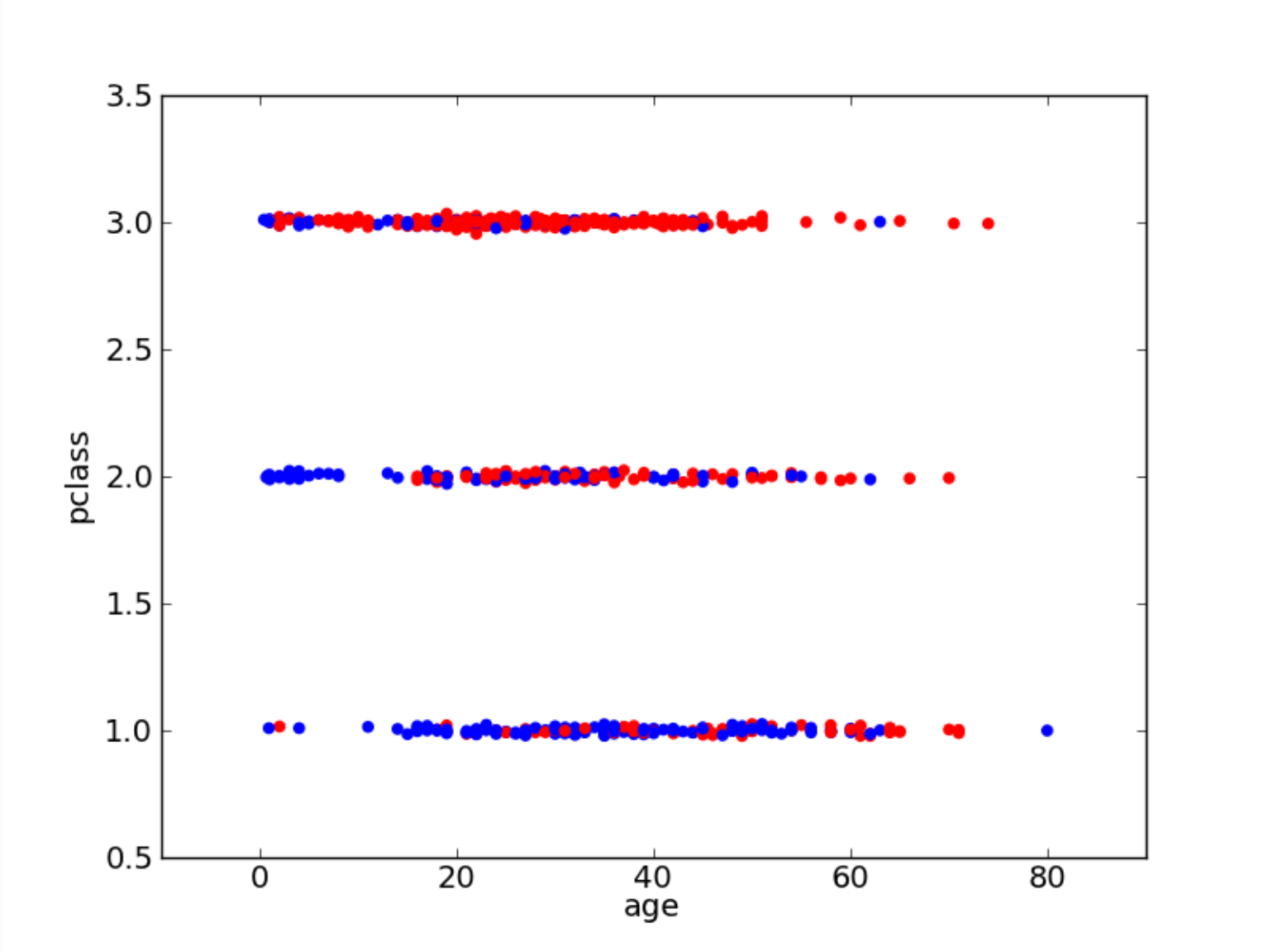
IF sex='female' THEN survive=yes
ELSE IF sex='male' THEN survive = no

```
confusion matrix

no   yes    <-- classified as
468  109  |   no
 81  233  |  yes
```

(468 + 233) / (468+109+81+233) = 79% correct (and 21% incorrect)

Not bad!


3.5
3.0
2.5
2.0
1.5
1.0
0.5
pclass
0
20
40
60
80
age

IF pclass=‘1’ THEN survive=yes
ELSE IF pclass=‘2’ THEN survive=yes
ELSE IF pclass=‘3’ THEN survive=no

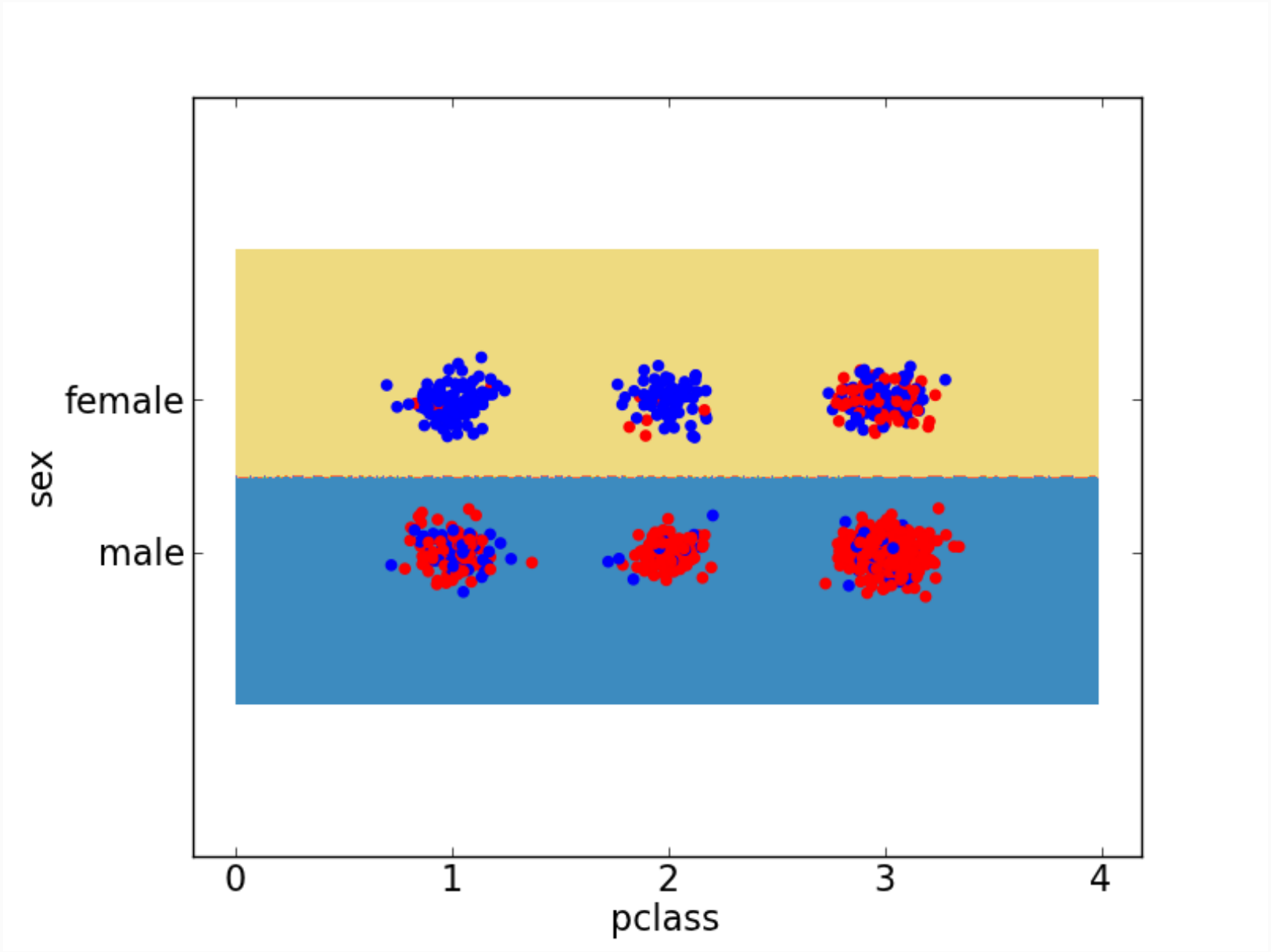
```
confusion matrix

no   yes    <-- classified as
372  119  |   no
177  223  |  yes
```

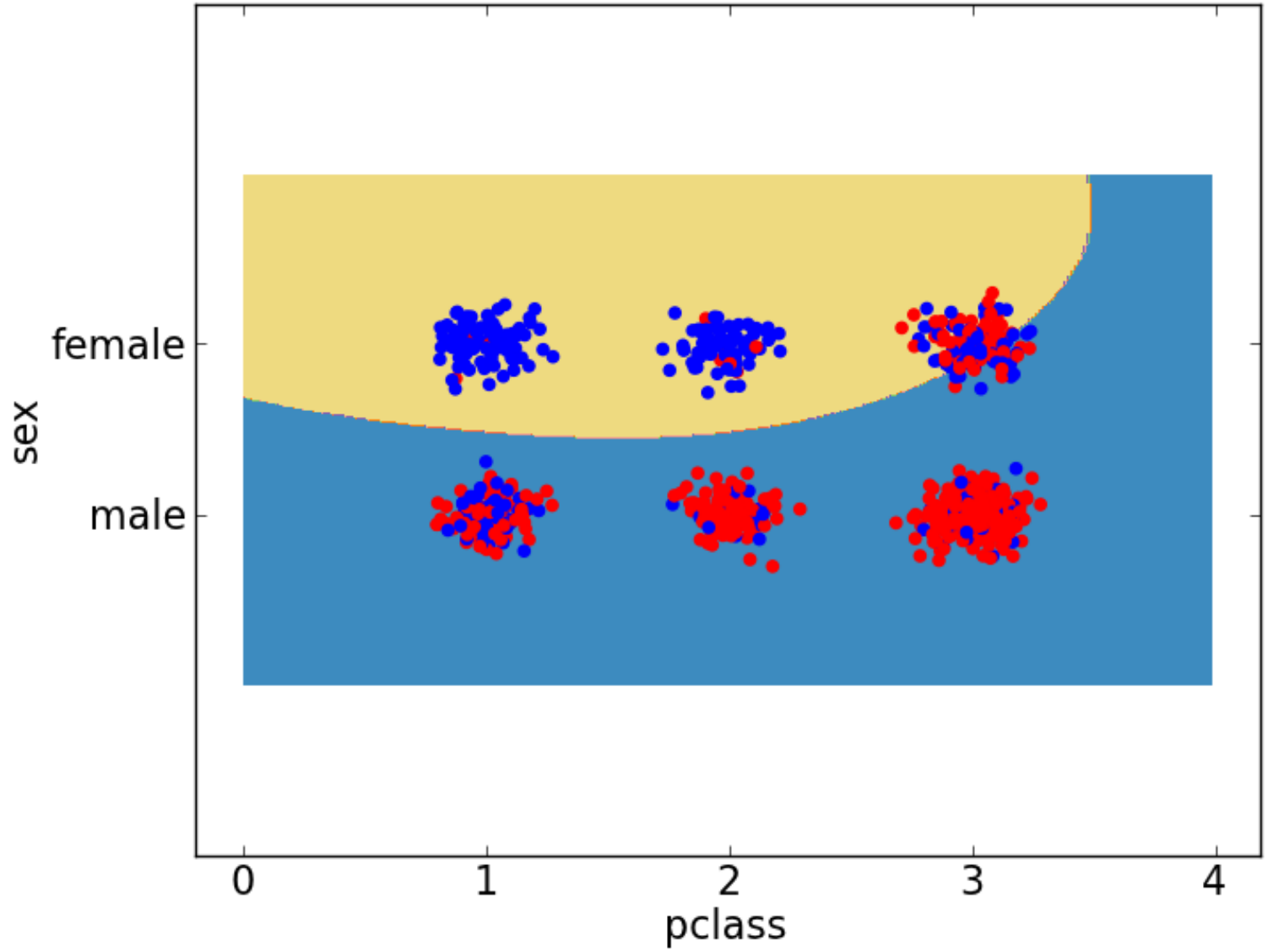
(372 + 223) / (372+119+223+177) = 67% correct (and 33% incorrect)

a little worse

Support Vector Machine Model, Titanic Data, Linear Kernel

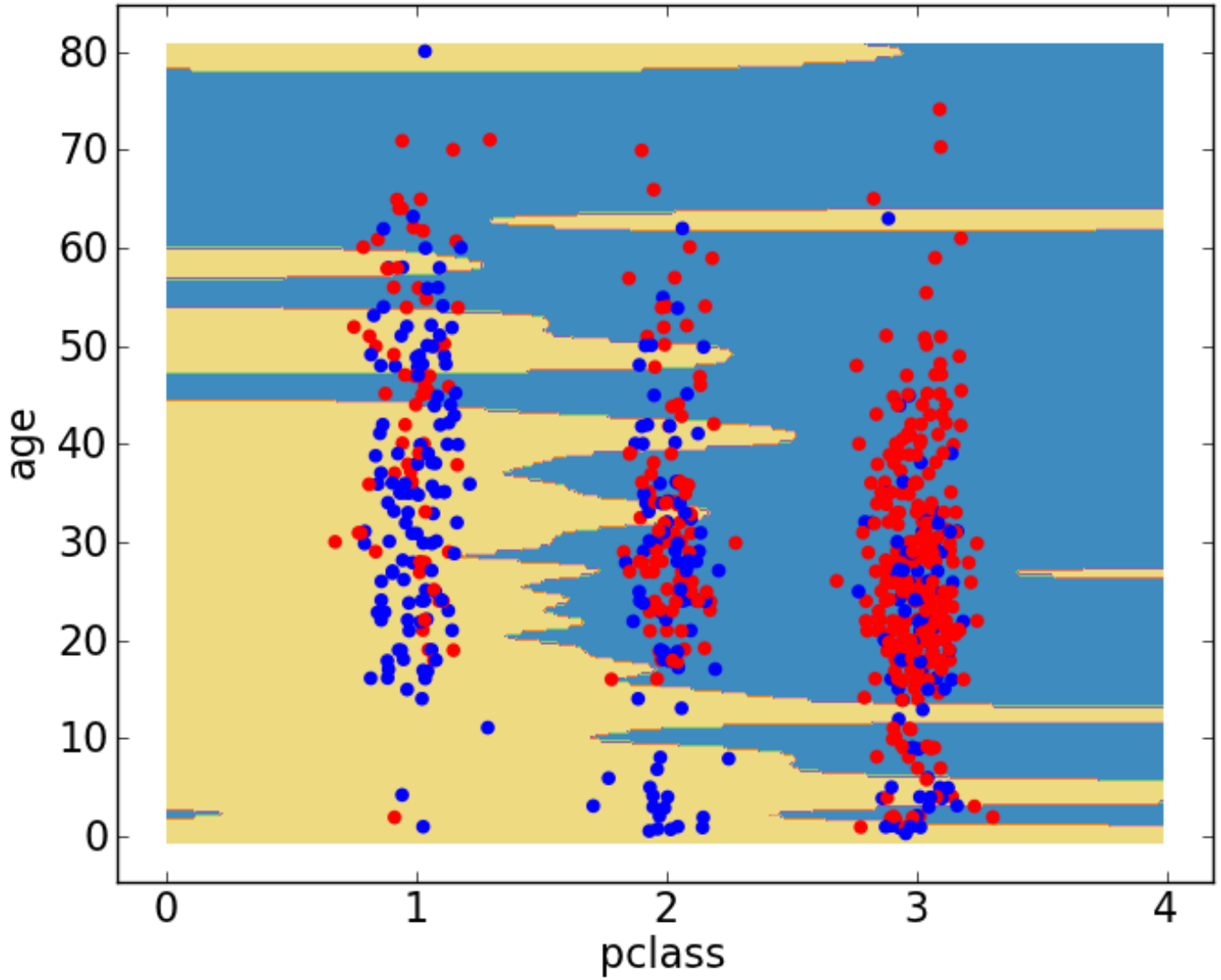


Support Vector Machine Model, RBF Kernel
Titanic Data



Support Vector Machine Model, RBF Kernel

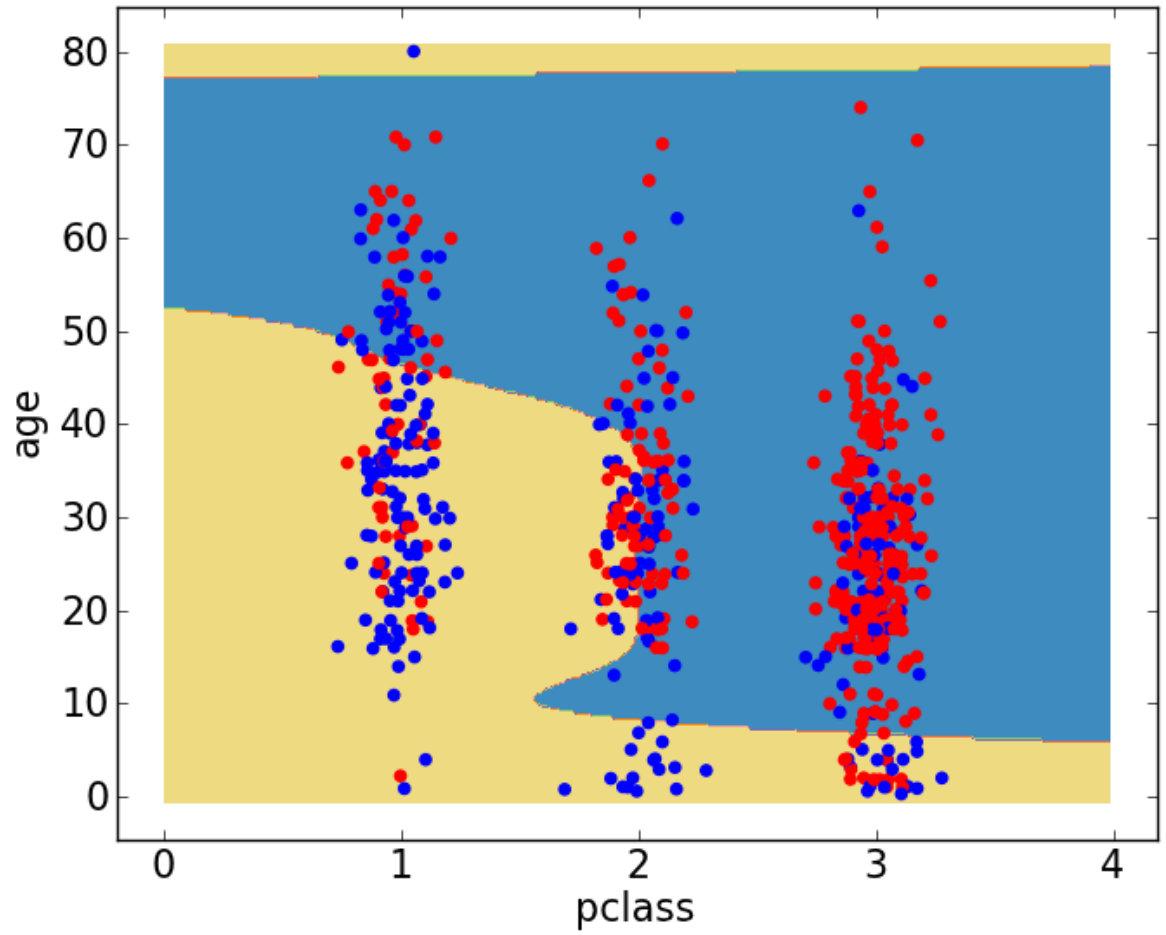
Titanic Data



overfitting?

Support Vector Machine Model, RBF Kernel
Titanic Data

A gamma, parameter that controls/balances model complexity against accuracy



Sparse Data

- SVM algorithms speed up dramatically if the data is sparse (i.e. many values are 0)
- Why? Because they compute lots and lots of dot products
- Sparse data compute dot products very efficiently
- Iterate only over nonzero values
- SVMs can process sparse datasets with 10,000s of attributes

Doing multi-class classification

- SVMs can only handle two-class outputs (i.e. a categorical output variable with arity 2).
- What can be done?
- Answer: with output arity N, learn N SVM's
  - SVM 1 learns "Output==1" vs "Output != 1"
  - SVM 2 learns "Output==2" vs "Output != 2"

    ….
  - SVM N learns "Output==N" vs "Output != N"
- Then to predict the output for a new input, just predict with each SVM and find out which one puts the prediction the furthest into the positive region.

What you need to know…

- First, try logistic regression. Easy, fast, stable. No 'tuning' parameters.
- Or try kNN classifier. It is simple ad fast!
- Equivalently, you can first try linear SVMs, but you need to tune 'C'
- If results are "good enough", stop
- Else, try SVMs with Gaussian kernels (RBF)

  Need to tune bandwidth, C – by using validation data…

Summary: Steps for Classification

- SVMs require vector of real numbers
  - Categorical variables → numeric data {R,G,B} → {0,0,1},…{1,0,0}
  - Scaling to the range [-1, +1] or [0,1]
- Select the kernel function to use
  - RBF is a reasonable first choice, two parameters C and γ
  - Grid search to identify best values for parameters
  - Use v-fold cross validation to ensure good performance on test data
- Unseen data can be classified using support vectors

Conclusion

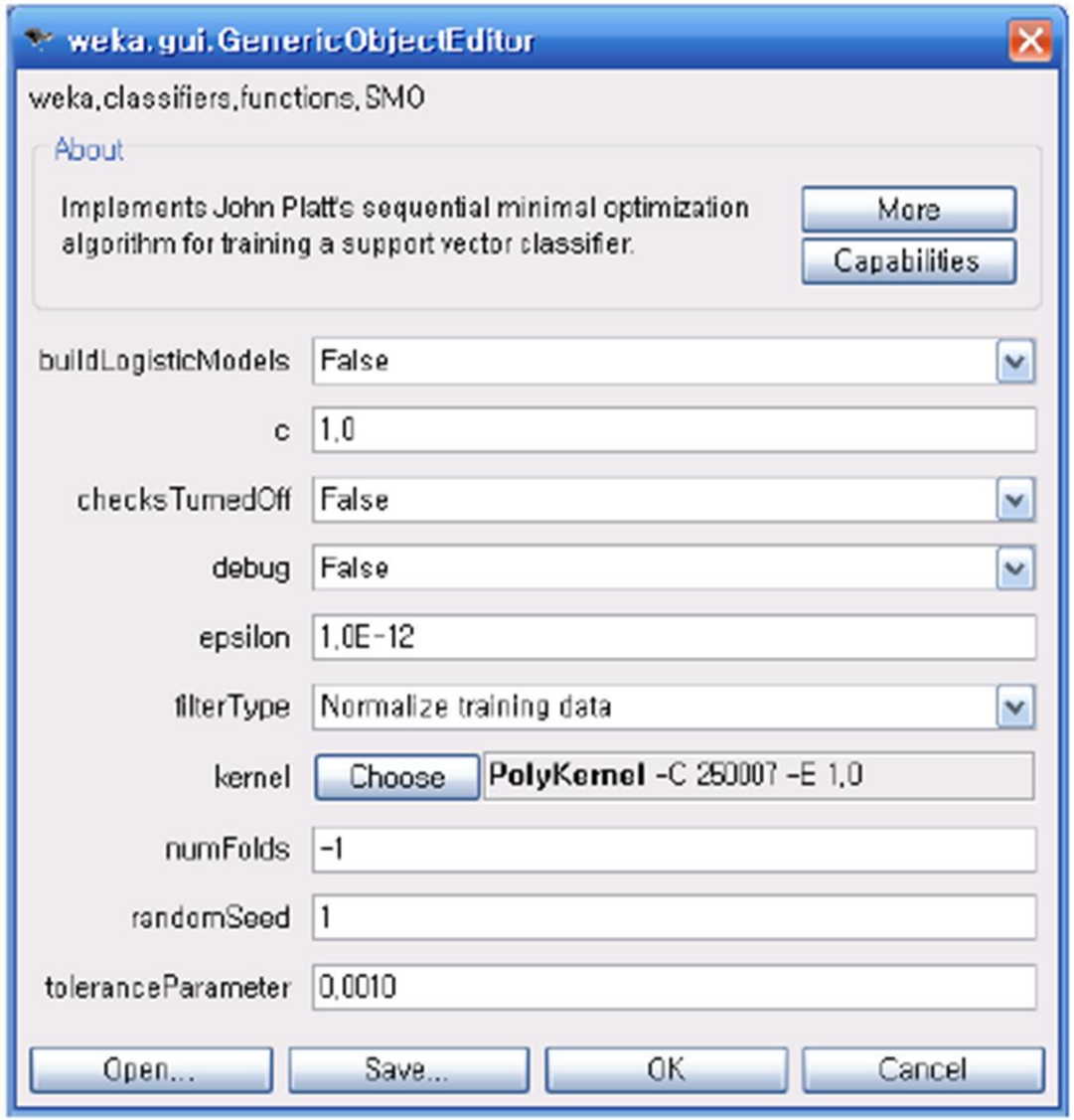
- SVMs balance between correctness and generalization
  - Decision boundaries
  - Margins
  - Support vector
- Two key concepts of SVM: maximize the margin and the kernel trick
- Many SVM implementations are available on the web for you to try on your data set!

Software

- A list of SVM implementation can be found at http://www.kernel-machines.org/software.html
- Some implementation (such as LIBSVM) can handle multi-class classification
- SVMLight is among one of the earliest implementation of SVM
- Several Matlab toolboxes for SVM are also available

In WEKA …

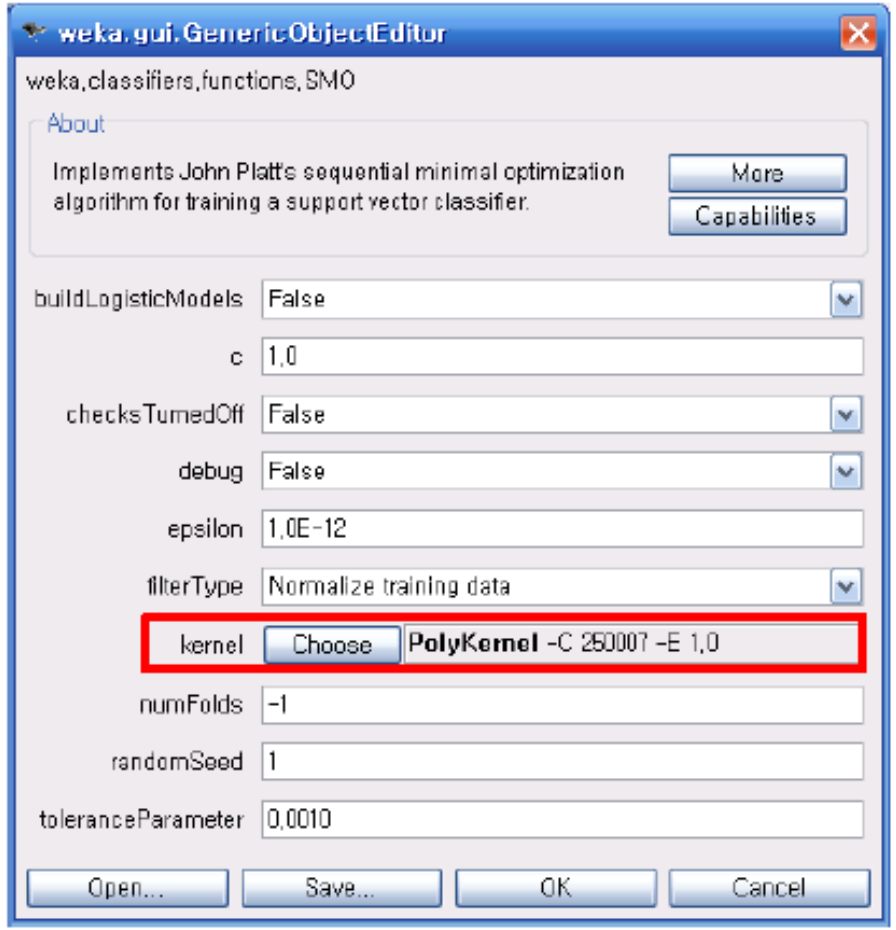
- Sequential minimal optimization (SMO) algorithm support vector *classification*
  - *weka.classifiers.functions.SMO*
- A Library for Support Vector Machines (libSVM)
  - *weka.classifiers.functions.libSVM*
- Sequential minimal optimization (SMO) algorithm support vector *regression*
  - *weka.classifiers.functions.SMOreg*



- buildLogisticModels: Whether to fit logistic models to the outputs (for proper probability estimates)
- numFolds: The number of folds for cross-validation used to generate training data for logistic models
- randomSeed: Random number seed for the cross-validation

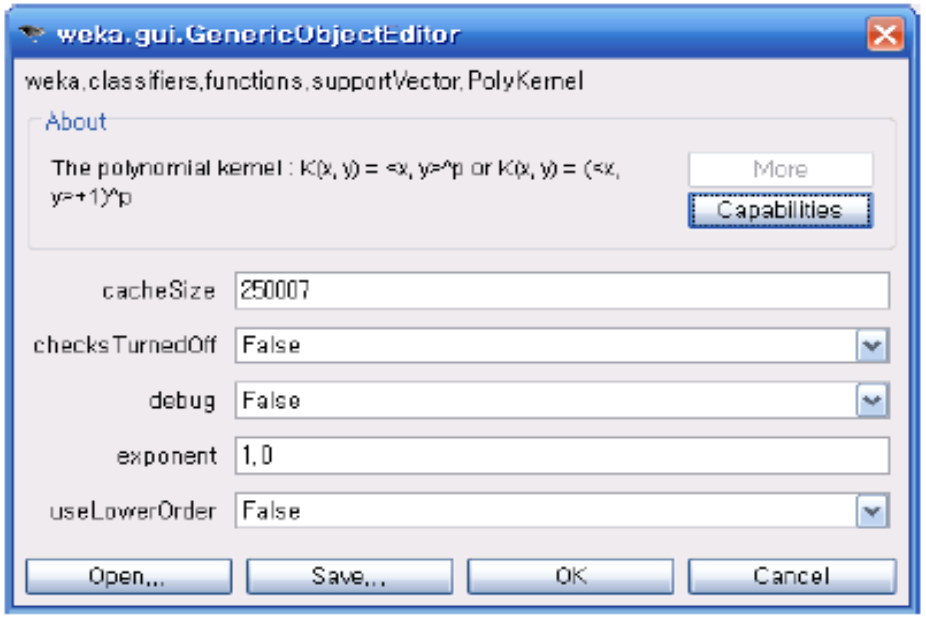
- c -- The complexity parameter C. It is the upper bound of alpha's
- filterType -- Determines how/if the data will be transformed.
- kernel -- The kernel to use.

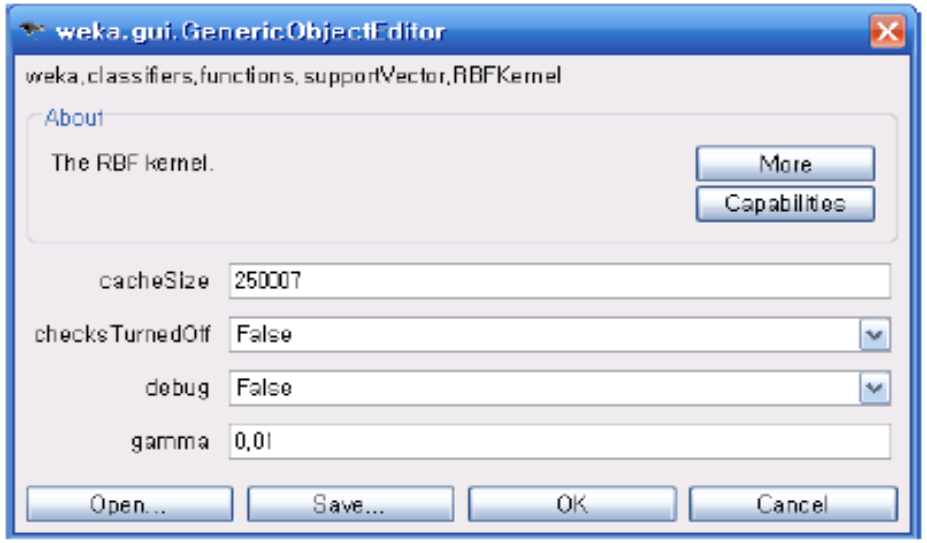
Parameter Setting Guide



- **Suggested**
  - buildLogisticModels: True
  - numFolds: 3 or 5
  - randomSeed: any value
  - C (complexity parameter, upper bound of alpha)
- **Your own choice**
  - filterType
  - Kernel and its subsequent parameters
  - Debug – if on, you can see intermediate results
- **Do not change!**
  - epsilon
  - checksTurnedOff
  - toleranceParameter

Parameter Setting Guide – Kernel



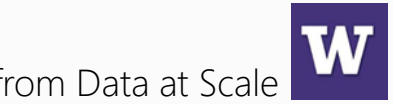


- **PolyKernel**
  - Try various exponents
    - Floating points are allowed
  - 1.0: linear kernel

- **RBFKernel**
  - Try various gammas
  - Gamma value corresponds to the inverse of the variance (width of the kernel)

5 Minute Break…

Data Exploration



Quick Hands On…

Use Weka to build a predictive model over the bank data set, select the simple decision tree J4.8 with default parameters

- Use **Bank Data A** for one model;
- Use **Bank Data B** for the second model;

Q. Is there a difference in the model performance between the two?

Q. If there is a performance difference, can you identify the reason?

Why Data Preprocessing?

Data in the real world is dirty

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- noisy: containing errors or outliers
- inconsistent: containing discrepancies in codes or names

No quality data, no quality modeling results!

- Quality decisions must be based on quality data
- Requires the collection and maintenance of gold standard data sets.

Data Understanding: Quantity

- Number of instances (records, objects)
  - *Rule of thumb: 5,000 or more desired*
  - If less, results are less reliable; use special methods (boosting, …)
- Number of attributes (fields)
  - *Rule of thumb: for each attribute, 10 or more instances*
  - If more fields, use feature reduction and selection
- Number of targets
  - *Rule of thumb: >100 for each class*
  - If very unbalanced, use stratified sampling or SMOTE

Major Tasks in Data Preprocessing

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data transformation

- Normalization and aggregation

Data reduction

- Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

- Data reduction but with particular importance, esp. for numerical data

Data Preparation

Data Transformation

Simple transformations can often have a large impact in performance

Example transformations (not all for performance improvement):

- Difference of two date attributes, distance between coordinate,…
- Ratio of two numeric (ratioscale) attributes, average for smoothing,….
- Concatenating the values of nominal attributes
- Encoding (probabilistic) cluster membership
- Adding noise to data (for robustness tests)
- Removing data randomly or selectively
- Obfuscating the data (for anonymity)

Intuition: add features that increase class discrimination (E, IG)…

Data Exploration Process

For each attribute:

Look at data summaries

- Identify potential problems and decide if an action needs to be taken

Visualize the distribution

- Identify potential problems (e.g., one dominant attribute value, even distribution, a relatively small number of unique values, distinct roughly equal to unique, outliers, etc.)
- Evaluate the potential usefulness of attributes

Feature **Selection**, *starts with you…*

Remove fields with no or little variability

- Examine the number of distinct field values
  - *Rule of thumb: remove a field where almost all values are the same*
- Irrelevant attributes in the input data will likely decrease the classification performance (supervised approaches)
- Your goal is to find the **smallest subset of attributes** leading to a higher classification accuracy than all attributes

Data Transformation

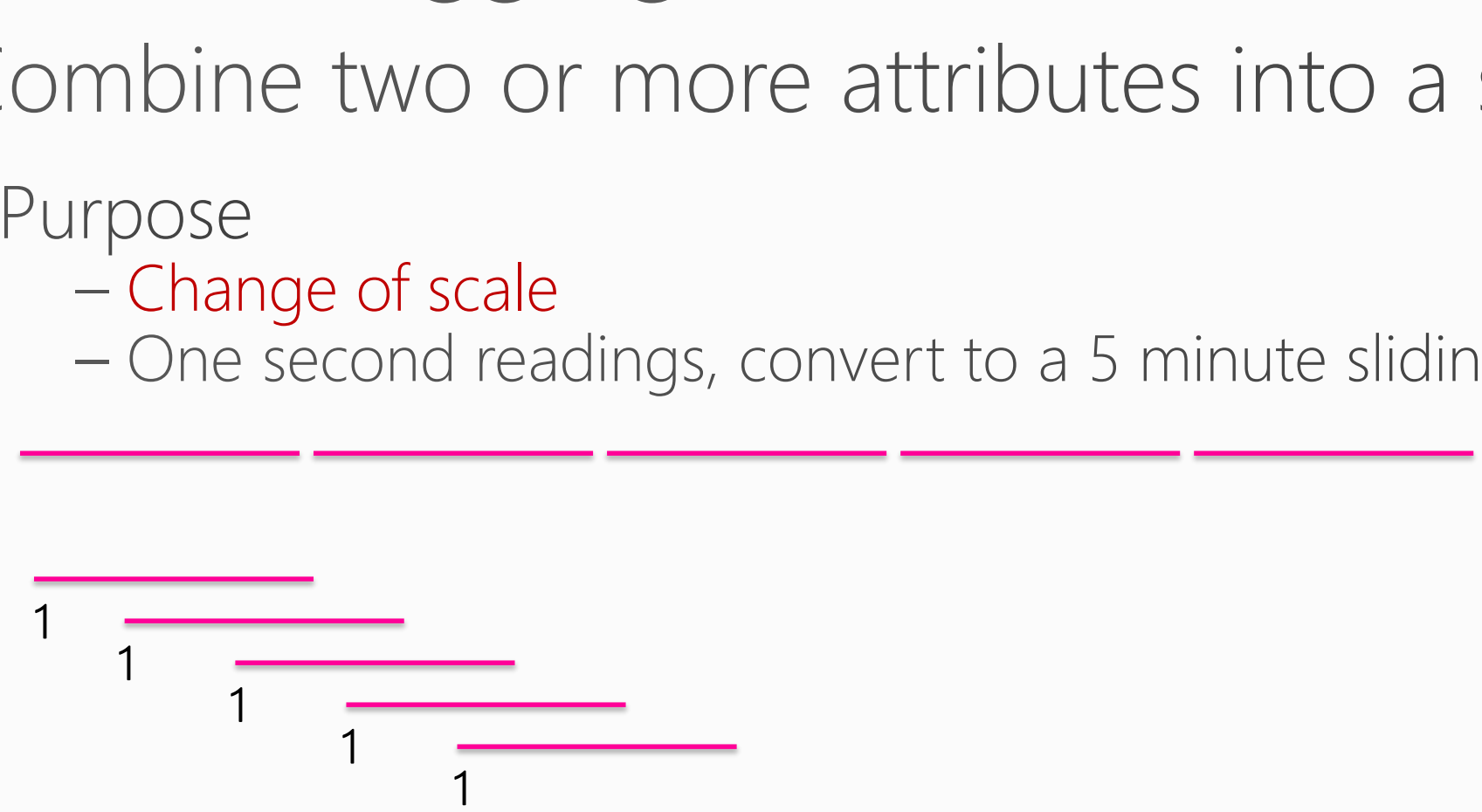
Why transform data?

- **Combine attributes**. For example, the ratio of two attributes might be more useful than keeping them separate
- **Normalizing data**. Having attributes on the same approximate scale helps many data mining algorithms (hence better models)
- **Simplifying data**. For example, working with discrete data is often more intuitive and helps the algorithms (hence better models)

Attribute Aggregation

Combine two or more attributes into a single attribute

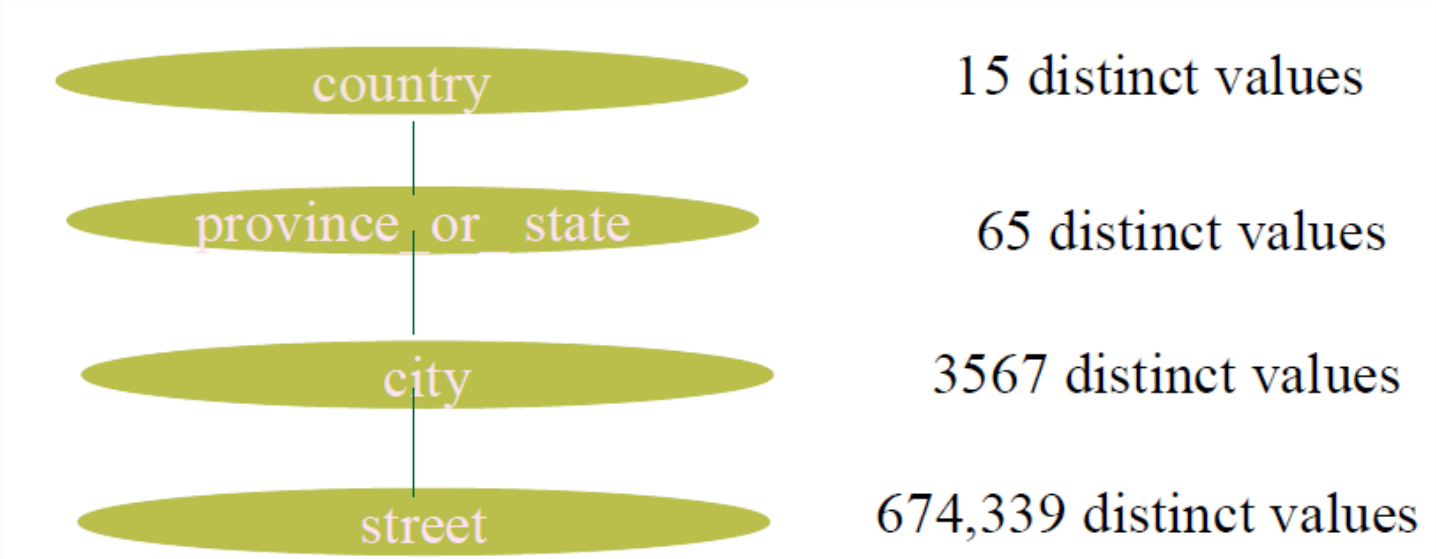
- Purpose
  - Change of scale
  - One second readings, convert to a 5 minute sliding/hopping window



Attribute Aggregation

Combine two or more attributes into a single attribute

- Purpose
  - Change of scale
  - One second readings, convert to a 5 minute sliding/hopping window
- Cities aggregated into regions, states, countries, etc
  - More "stable" data



Attribute Aggregation

Combine two or more attributes into a single attribute

- Purpose
  - Change of scale
  - One second readings, convert to a 5 minute sliding/hopping window
- Cities aggregated into regions, states, countries, etc
  - More "stable" data
- Ratio of two attributes carries more information that separate values
  - More "information" for prediction

Aggregated data tends to have less variability and potentially more information for better predictions…

Explore the Weka Filters

Weka has many filters helpful in preprocessing the data

- Attribute filters
  - Add, remove, or transform attributes
- Instance filters
  - Add, remove, or transform instances

- Process
  - Choose for drop-down menu
  - Edit parameters (if any)
  - Apply

Data Cleaning

- Missing values
  - Weka reports % of missing values
  - Can use filter called ReplaceMissingValues
- Noisy data
  - Due to uncertainty or errors
  - Weka reports unique values
  - Useful filters include
    - RemoveMisclassified
    - MergeTwoValues

Explore the Weka Filters

The data transformation filters in Weka include:

- Add – adds a new attribute to a data set
- AddExpression – adds a new attribute, value computed by expr
- MakeIndicator – filter that creates a new data set, replacing nominal variable with Boolean variable and values;
- NumericTransform – transforms numeric attributes given a transform method: abs, max, min, mean, mode;
- Normalize – by default [0…1], [-1…1], other specified ranges possible;
- Standardize – standardizes all numeric attributes in the data set to have zero means and unit variance;

Discretization

Discretization reduces number of values for a continuous attribute

- Why?
  - Some methods can only use nominal data
    - E.g., in Weka ID3 and Apriori algorithms
  - Helpful if data needs to be sorted frequently (e.g., when constructing a decision tree)

Unsupervised Discretization, Binning

Unsupervised - does not account for classes

- Equal-interval binning – the simplest unsupervised discretization method, determines the minimum and maximum values of the discretized attribute and then divides the range into the user-defined number of equal width discrete intervals.

- Equal-frequency binning – the unsupervised method, which divides the sorted values into k intervals so that each interval contains approximately the same number of training instances. Thus each interval contains n / k (possibly duplicated) adjacent values. k is a user predefined parameter.

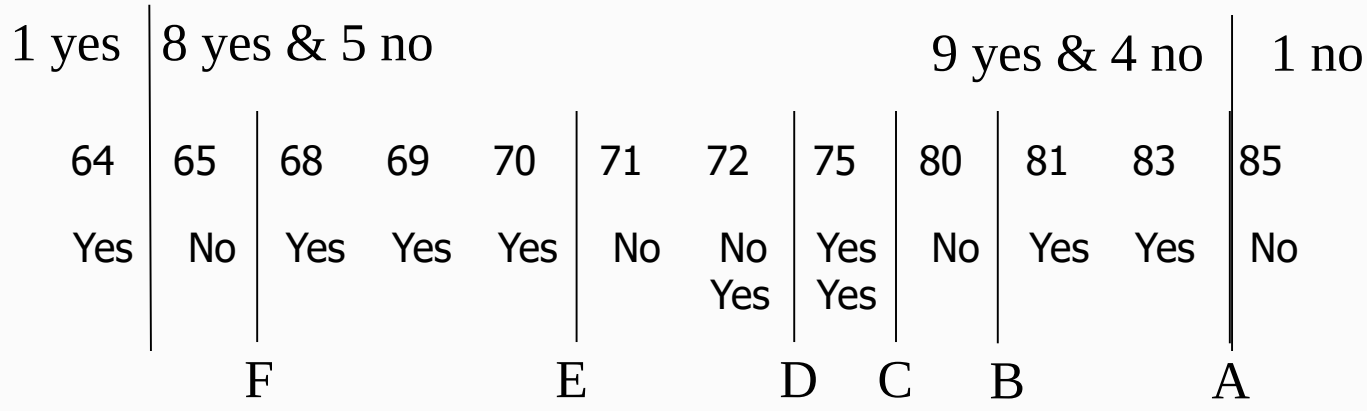
Binning, cont'd

Sort data in ascending order: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal-depth bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Supervised Discretization

- Take classification into account
- Use “entropy” to measure information gain
- Goal: Discretise into 'pure' intervals
- Usually no way to get completely pure intervals:



Explore the WEKA features, use in class next week

Discretize

- \Weka\diabetes
- Discretize "age" (equal bins vs equal frequency)

NumericToNominal

- \Weka\diabetes
- Discretize "age" (vs "Discretize" method)

NominalToBinary

- \UCI\autos
- Convert "num-of-doors"
- Convert "drive-wheels"

Data Reduction

- Another way is to reduce the size of the data before applying a learning algorithm (preprocessing)
- Some strategies
  - Sampling
  - Dimensionality reduction
  - Data compression
  - Numerosity reduction

Sampling Principle

Key principle for effective sampling

- Using a sample will work almost as well as using the entire data set, if the sample is representative
- A sample is representative if it has approximately the same property (of interest) as the original set of data

Sampling Techniques

- Different samples
  - Sample without replacement
  - Sample with replacement
  - Cluster sample
  - *Stratified sample*
- Complexity of sampling actually *sublinear*, that is, the complexity is O(*s*) where *s* is the number of samples and *s*<<*n*

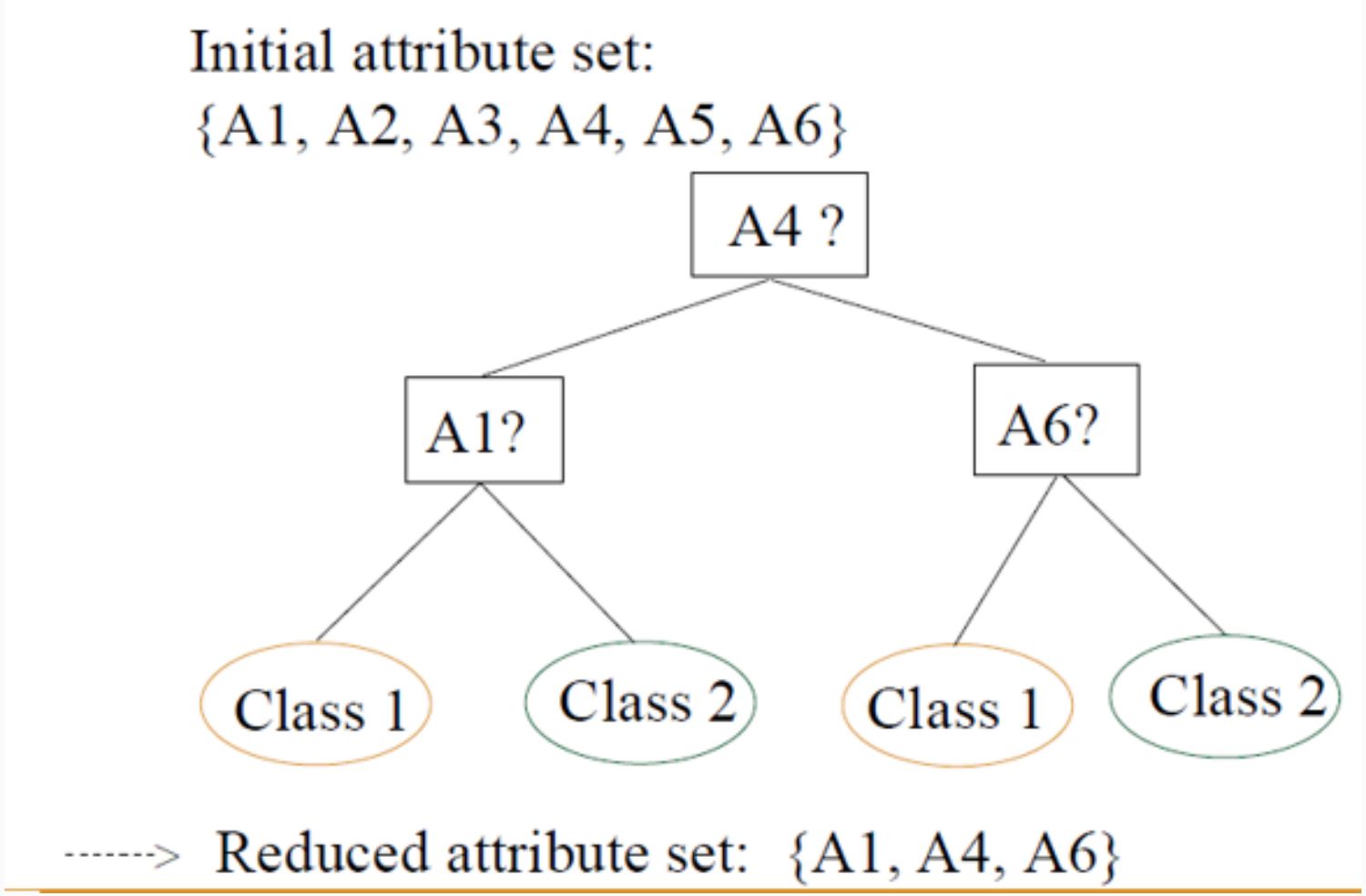
Dimensionality Reduction

- Remove irrelevant, weakly relevant, and redundant attributes
- Attribute selection
  - Many automated methods available
  - E.g., forward selection, backwards elimination, genetic algorithm search
- Often results in a much smaller learning problem
- Often little degeneration in predictive performance or even better performance

Dimensionality Reduction

One we have used in previous exercises?...

Decision Tree Induction...



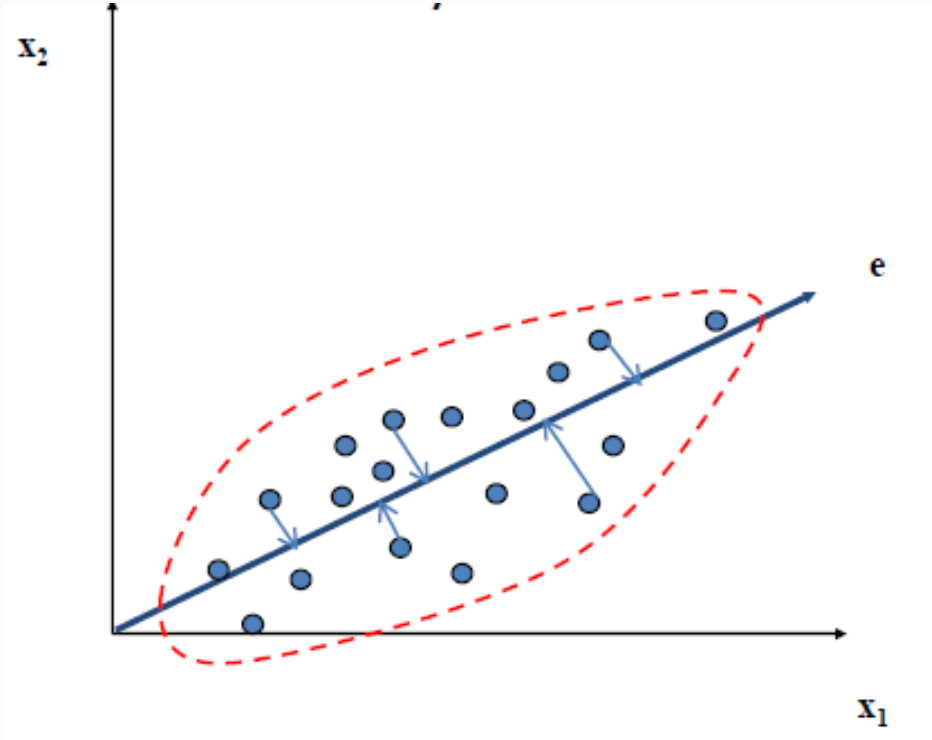
Data Compression

Transform the data into a smaller space

- *Principle Component Analysis*
- Canonical Correlation Analysis (CCA)
- Linear Discriminant Analysis (LDA)
- Independent Component Analysis (ICA)
- Manifold Learning

Principal Component Analysis (PCA)

- Find a projection that captures largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.



Principal Component Analysis (Steps)

Given data from *n*-dimensions (*n* features), *find k ≤ n new features (principal components) that can best represent data*

- Normalize input data: each feature falls within the same range
- Compute *k principal components* (details omitted)
- Each input data is projected in the new *k*-dimensional space
- The new features (principal components) are sorted in order of decreasing "significance" or strength
- Eliminate *weak components / features* to reduce dimensionality.

→ *Works for numeric data only…*

PCA Exploration in WEKA, try it out…

\UCI\breast-w

- Accuracy with all features
- PrincipalComponents (data transformation)
- Visualize/save transformed data (first two features, last two features)
- Accuracy with all transformed features
- Accuracy with top 1 or 2 feature(s)

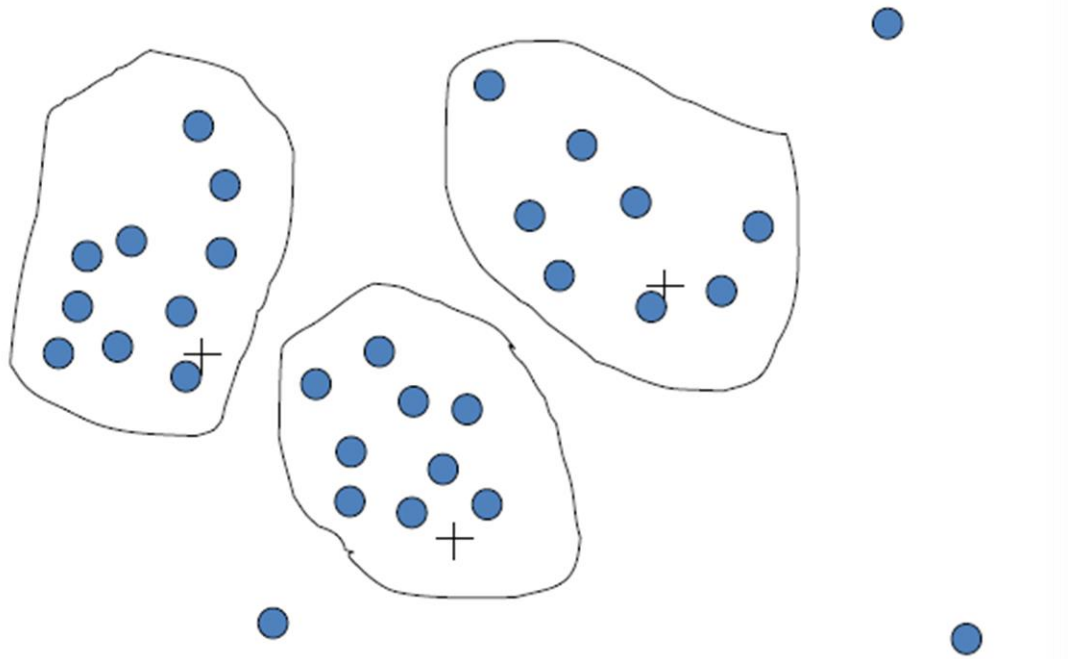
PrincipalComponents under Attribute Selection tab

Numerosity Reduction

- Clustering
  - Data objects (instance) that are in the same cluster can be treated as the same instance
  - Must use a scalable clustering algorithm
- Sampling
  - Randomly select a subset of the instances to be used

Clustering

- Partition data set into clusters, and one can store cluster representation only
- Can be very effective if data is clustered but not if data is "smeared"



5 minute Break…

Attribute Selection

Reasons for Attribute Selection

Simpler model

- More transparent
- Easier to interpret

Faster model induction

- What about overall time?

Structural knowledge

- Knowing which attributes are important may be inherently important to the application

What about the accuracy?

Variable Ranking – Feature Selection

A simple method for feature selection using variable ranking is to select the k highest ranked features according to S.

This is usually not optimal

but often preferable to other, more complicated methods

Computationally efficient(!): only calculation and sorting of n scores

Feature Ranking

Steps

1. Rank all the individual features according to certain criteria (e.g., information gain, gain ratio, $\chi 2$)
2. Select / keep top *N* features

Properties

- Usually independent of the learning algorithm to be used
- Efficient (no search process)
- Hard to determine the threshold
- Unable to consider correlation between features

By convention a high score indicates a valuable feature.

Ranking Criteria – Correlation

Correlation Criteria:

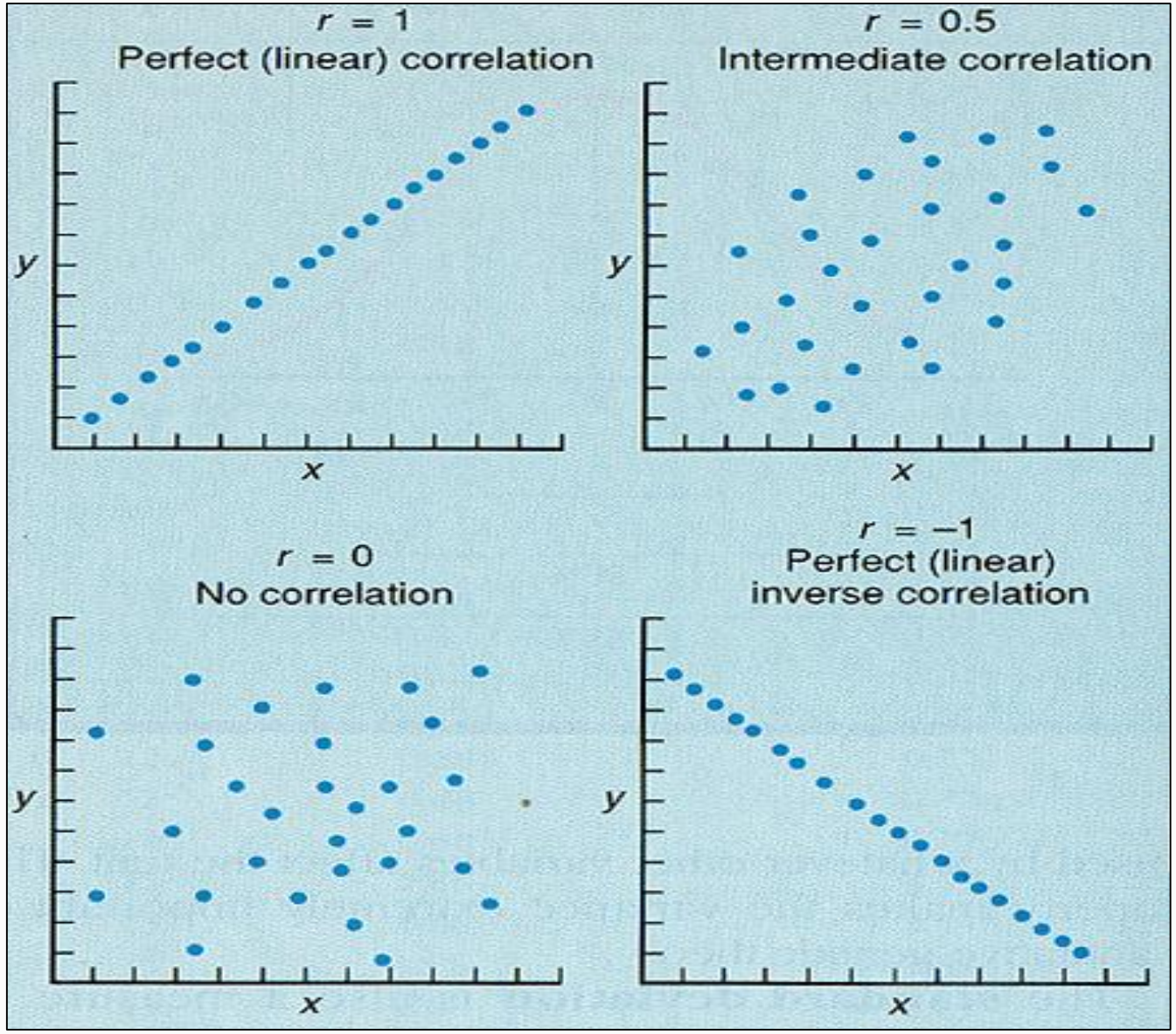
- Pearson correlation coefficient

- mostly $R(x_i,y)^2$ or $|R(x_i,y)|$ is used

- measure for the goodness of **linear** fit of x_i and y.
  (can only detect linear dependencies between variable and target.)

The higher the correlation between the feature and target, the higher the score!

Ranking Criteria – Correlation



Ranking Criteria – Correlation

Correlation Criteria:

- $R(X,Y) \in [-1,1]$
- mostly $R(x_i,y)^2$ or $|R(x_i,y)|$ is used
- measure for the goodness of **linear** fit of x_i and y.
  (can only detect linear dependencies between variable and target.)

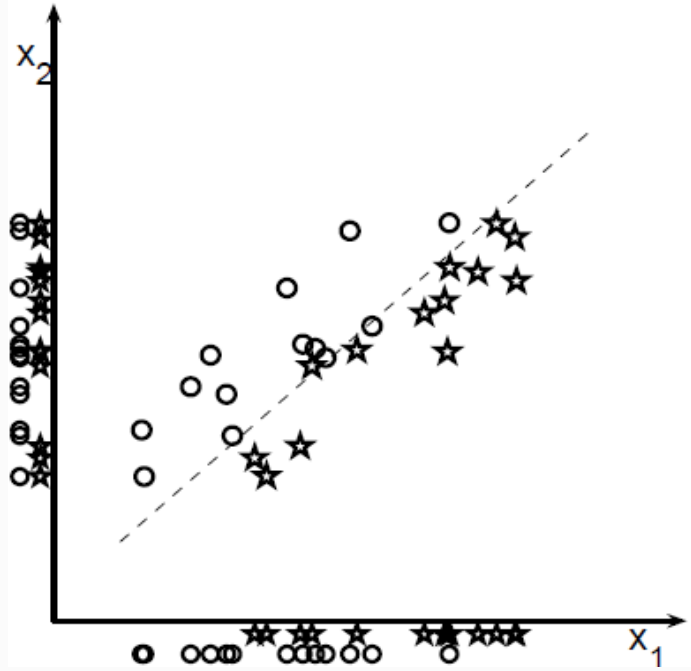
Ranking Criteria – Correlation

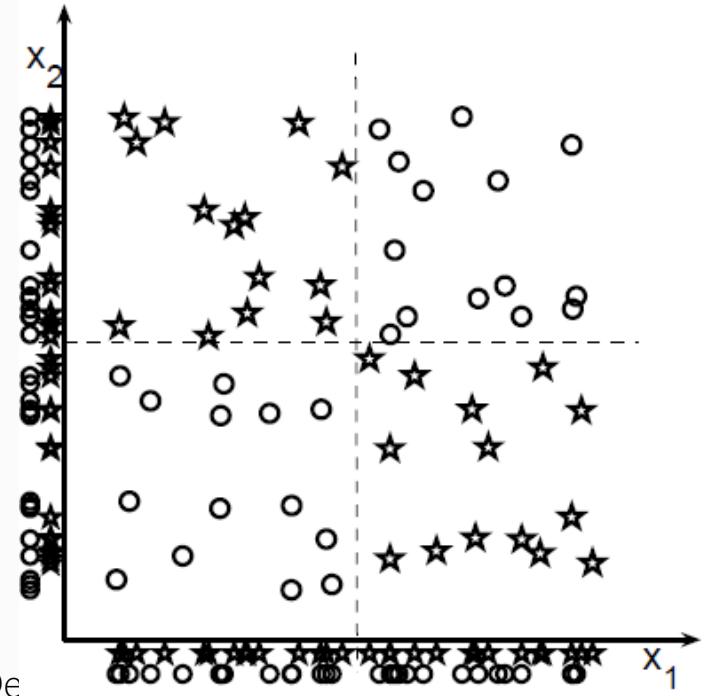
Questions:

- Can variables with small score be automatically discarded ?
- Can a useless variable (i.e. one with a small score) be useful together with others ?
- Can two variables that are useless by themselves can be useful together?)

Ranking Criteria – Correlation
Multivariate Dependencies

- Can variables with small score be discarded without further consideration?
  NO!
- Even variables with small score can improve class separability!
- Multivariate dependencies
  (a) Feature X2 is individually irrelevant to Y, but it becomes relevant in the context of feature X1;
  (b) Two individually irrelevant features become relevant when taken jointly





De

Ranking Criteria – Correlation

- Correlation between variables and target are not enough to assess relevance!
- Correlation / covariance between pairs of variables has to be considered too! (potentially very difficult to perform)
- **Diversity** of features

Feature Selection and Engineering

Filter Methods

Results in either

- Ranked list of attributes
  - Typical when each attribute is evaluated individually
  - Must select how many to keep
- A selected subset of attributes
  - Forward selection
  - Best first
  - Random search such as genetic algorithm

Filter Evaluation Examples

- Information Gain
- Gain ratio
- Relief

Optimal Correlation

- High correlation with class attribute
- Low correlation with other attributes

Feature Selection and Engineering

Wrapper Methods

Wrapper methods find features that work best with some particular learning algorithm:

- Best features for k-Means, SVM or a neural network may not be best features for decision trees;
- Can eliminate features learning algorithm "has trouble with";

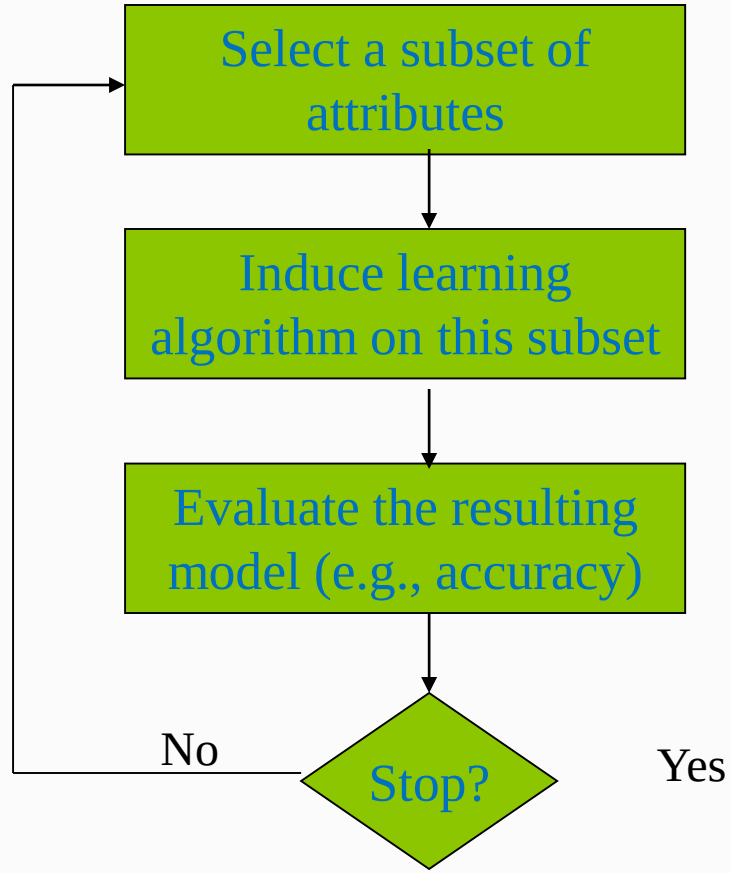
Forward stepwise selection, *forward reference…*

Backwards elimination, *backward reference…*

Bi-directional stepwise selection and elimination

Wrappers

- "Wrap around" the learning algorithm
- Must therefore always evaluate subsets
- Return the best subset of attributes
- Apply for each learning algorithm
- Use same search methods as before



Forward Feature Selection

Steps

1. First select the best single-feature (according to the learning algorithm)
2. Repeat (until some stop criterion is met):
   Select the next best feature, given the already picked features

Properties

- Usually learning algorithm dependent
- Feature correlation is considered
- More reliable
- Inefficient

Feature Selection and Engineering

Heuristics

SFS (Sequential Forward Selection)

- Begins with zero attributes
- Evaluates all features subsets w/ exactly 1 feature
- Selects the one with the best performance
- Adds to this subsets the feature that yields the best performance for subsets of next larger size
- If **EVAL()** is a heuristics measure, the feature selection algorithm acts as a filter, extracting features to be used by the main algorithm; If it is the actual accuracy, it acts as a wrapper around that algorithm

Feature Selection and Engineering

Heuristics

SFS (Sequential Forward Selection)

```
SS = 0
BestEval = 0
REPEAT
   BestF = None
   FOR each feature F in FS AND NOT in SS
      SS' = SS ∪ {F}
      IF Eval(SS') > BestEval THEN
                BestF = F; BestEval = Eval(SS')
   IF BestF <> None THEN SS = SS ∪ {BestF}
UNTIL BestF = None OR SS = FS
RETURN SS
```

Backward Feature Elimination

Steps

1. First build a model based on all the features
2. Repeat (until some criterion is met):
   Eliminate the feature that makes the least contribution.

Properties

- Usually learning algorithm dependent
- Feature correlation is considered
- More reliable
- Inefficient

Feature Selection and Engineering

Heuristics

SBS (Sequential Backward Selection)

```
SS =FS
BestEval = Eval(SS)
REPEAT
   WorstF = None
   FOR each feature in F in FS
      SS' = SS - {F}
      IF Eval(SS') >= BestEval THEN
               WorstF = F; BestEval = Eval(SS')
   IF WorstF <> None THEN SS = SS - {WorstF}
UNTIL WorstF = None OR SS = 0
RETURN SS
```

Weka, try it yourself (will cover in class next week)

Feature Ranking

- \Weka\weather
- ChiSquared, InfoGain, GainRatio

FFS & BFE

- \Weka\Diabetes
- ClassifierSubsetEval + GreedyStepwise

Summary, Heuristic Search in Feature Selection

- Given d features, there are 2d possible feature combinations
  - Exhaust search won't work
  - Heuristics have to be applied
- Typical heuristic feature selection methods:
  - Feature ranking
  - Forward feature selection
  - Backward feature elimination
  - Bidirectional search (selection + elimination)
  - Search based on evolution algorithm
  - ……

Summary

- In real world applications, data preprocessing usually occupies majority of the workload in a data mining task.
- Domain knowledge is usually required to do good data preprocessing.
- To improve a predictive performance of a model
  - Improve learning algorithms (different algorithms, different parameters)
- Most data mining research focuses on here
    - Improve data quality ---- data preprocessing
- Deserve more attention!

Filter vs Wrapper Model

Filter model

- –Separating feature selection from learning
- –Relying on general characteristics of data (information, etc.)
- –No bias toward any learning algorithm, fast
- –Feature ranking usually falls into here

Wrapper model

- –Relying on a predetermined learning algorithm
- –Using predictive accuracy as goodness measure
- –High accuracy, computationally expensive
- –FFS, BFE usually fall into here

Feature Selection and Engineering

Important points 1/2

- Feature selection can significantly increase the performance of a learning algorithm (both accuracy and computation time) – but it is not easy!

- Relevance <-> Optimality

- Correlation and Mutual information between single variables and the target are often used as Ranking-Criteria of variables.

Feature Selection and Engineering

Important points 2/2

- One cannot automatically discard variables with small scores – they may still be useful together with other variables.
- Filters – Wrappers - Embedded Methods
- How to search the space of all feature subsets ?
- How to asses performance of a learner that uses a particular feature subset ?

Course Project Discussion

Data Science

Deriving Knowledge from Data at Scale

That's all for tonight…