Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry
patrick@mcsharry.net
www.mcsharry.net
Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Course outline

Week	Description
1	Statistical learning
2	Linear models
3	Nonlinear models
4	Supervised learning
5	Unsupervised learning
6	Ensemble approaches

Applied Machine Learning

WEEK 8A

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Prediction using linear regression	10
2	Discussion	Regularization	10
3	Case study	Prostrate Cancer	10
4	Analysis	Parameter estimation in linear models	20
5	Matlab Demo	Techniques for linear regression	20
6	Q&A	Matlab questions and feedback	10

Review

- Number of observations: n
- Model complexity, parameters: p
- Noisy measurements
- Over-fitting
- Training and testing
- Curve fitting

Poll

- Which of the following situations is particularly challenging for model fitting?
  1. Low noise and few observations
  2. Low noise and many observations
  3. High noise and few observations
  4. High noise and many observations

- www.slido.com #97725

Data: response and predictors

- Consider the response (dependent variable):

$$\mathbf{y} = (y_1, \ldots, y_n)^T$$

and a n x p model matrix **X** defined by:

$$\mathbf{X} = [\mathbf{x}_1 \mid \ldots \mid \mathbf{x}_p]$$

containing predictors (features or explanatory variables):

$$\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T, \; j = 1, \ldots, p$$

Linear regression

- Given *p* predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$, the response **y** is predicted by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \cdots + \mathbf{x}_p\hat{\beta}_p$$

- A model fitting procedure produces the vector of parameters

$$\hat{\beta} = \left[\hat{\beta}_0, \cdots, \hat{\beta}_p\right]$$

Ordinary Least Squares

- We define the ordinary least squares criterion as:

$$L(\beta) = |y - X\beta|^2$$

- The ordinary least squares estimator is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta)$$

R^2 - coefficient of determination

- The coefficient of determination, R^2, measures the proportion of variability in a data set that is accounted for by a statistical model
- In the case of linear regression, we can decompose the sum of the squares into a part due to the regression and the residuals, such that $SS_{tot} = SS_{reg} + SS_{res}$ where

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2 \quad SS_{res} = \sum_i (y_i - \hat{y}_i)^2$$

- and R^2 is defined as

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- R^2 measures the amount of variance explained by the model given by the ratio of the explained variance (variance of the model's predictions) with the total variance (of the data)

R^2 and correlation

- Coefficient of determination, R^2 is related to the correlation coefficient
- Both attempt to quantify how well a linear model fits to a data set
- The further the points are scattered from the line, the smaller is the value of R^2
- R^2 is square of the coefficient of correlation which is typically symbolized by r

Mean Squared Error

- The mean-square-error is given by

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

- It represents a measure of forecast performance which is analogous to the least squares parameter estimation technique
- If the forecast errors are not normally distributed, MSE may give misleading results

Mean Absolute Error

- The mean absolute error is given by

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | \hat{y}_i - y_i |$$

- This forecast measure focuses on the magnitude of the errors
- It is more robust than MSE as the large errors are not squared
- It is commonly used in wind energy forecasting and may be given as a fraction of the total energy being generated

Quiz

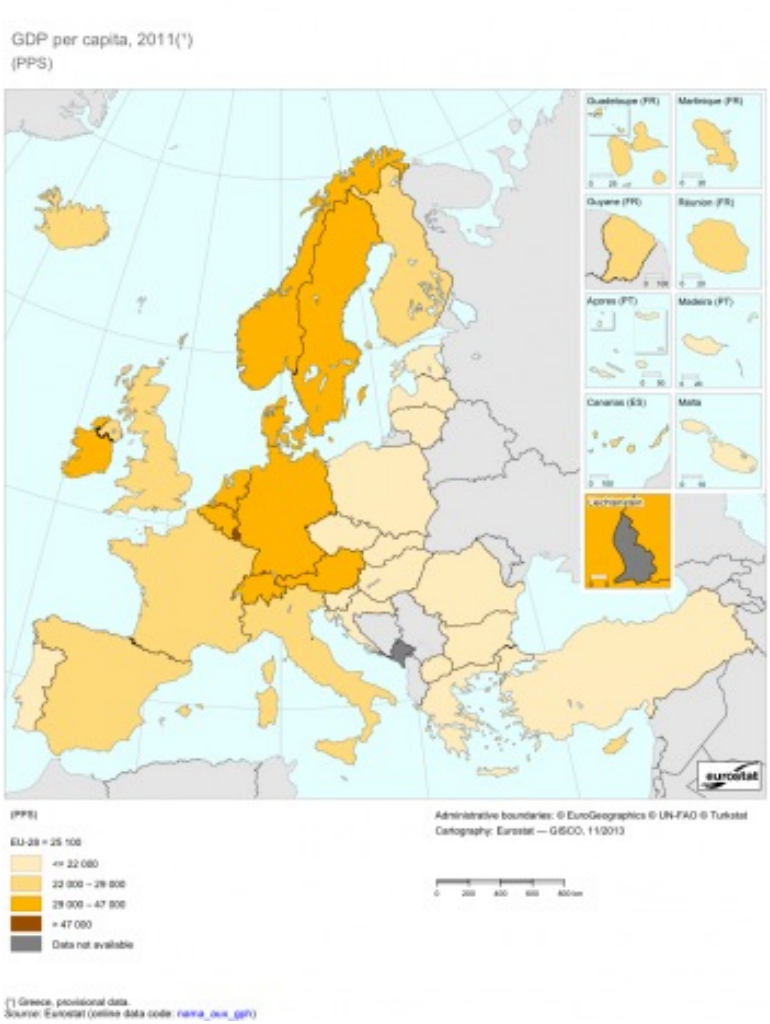
- When fitting a model to data, the objective is to maximise:
  1. Mean squared error
  2. Sum of the squared errors
  3. Mean absolute error
  4. Coefficient of determination

www.Slido.com #97725

Quality of life

Quality of life: 8+1 dimensions (Eurostat):

- Material living conditions (income, consumption and material conditions)
- Productive or main activity
- Health
- Education
- Leisure and social interactions
- Economic and physical safety
- Governance and basic rights
- Natural and living environment
- Overall experience of life



Source: http://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators_-_measuring_quality_of_life

Quality of Life variables

- Material well-being as measured by GDP per capita (in $, at 2006 constant PPPS)
- Life expectancy at birth
- The quality of family life based primarily on divorce rates
- The state of political freedoms
- Job security (measured by the unemployment rate)
- Climate (measured by two variables: the average deviation of minimum and maximum monthly temperatures from 14 degrees Celsius; and the number of months in the year with less than 30mm rainfall)
- Personal physical security ratings (based primarily on recorded homicide rates and ratings for risk from crime and terrorism)
- Quality of community life (based on membership in social organisations)
- Governance (measured by ratings for corruption)
- Gender equality (measured by the share of seats in parliament held by women)

Example: Quality of Life

Regression statistics	
Multiple R	0.919
Adjusted R square	0.823
Standard error	0.482
Observations	74

	Coefficients	*Standard error*	*Statistic*
Constant	2.7959	0.7890	3.5435
GDP per person	0.00003	0.00001	3.5247
Life expectancy	0.0448	0.0106	4.2299
Political freedom	-0.1052	0.0561	-1.8749
Job security	-0.0217	0.0099	-2.2062
Family life	-0.1878	0.0640	-2.9349
Climate and geography	-1.3534	0.4691	-2.8852
Political stability	0.1519	0.0520	2.9247
Gender equality	0.7423	0.5428	1.3676
Community life	0.3865	0.1237	3.1255

Linear regression links the results of subjective life-satisfaction surveys to the objective determinants of quality of life across countries.

Gauss-Markov Theorem

- Gauss–Markov theorem states that in a linear regression model in which the errors have expectation zero and are uncorrelated and have equal variances, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator.
- However, unbiased estimates may not be appropriate when we want a model that is capable of generalizing to new data!

Poll

- Generalization of a model relies on:

1. Variable selection
2. Parameter estimation
3. Avoiding overfitting
4. All of the above

www.Slido.com #97725

Model selection

- Given a collection of predictors, how should we select predictors for model construction?
- Stepwise regression is one approach.
- Forward selection and backward selection.
- This involves multiple model estimations while trying out different combinations of predictors.
- Can we combine model selection and model estimation?

Occam's razor

William of Occam studied theology at the University of Oxford from 1309 to 1321, but never completed his master's degree

- Occam's razor (principle of parsimony): seek the simplest model that describes the data
- Aim to select a model that can forecast future activity by avoiding over-fitting problems
- A nonlinear model may be more parsimonious than a linear model
- How complicated should a model be?
- How can we evaluate and compare models?
- What are the appropriate benchmarks?

Errors

- Suppose the data generating process is G.
- Noisy measurements are observed:

  $y_n = G(x_n) + \varepsilon_n$ with $\varepsilon_n \sim N(0,\sigma^2)$
- Estimate a model F using training data D.
- The error resulting from using this model F to predict an unseen input x with response y is

  $y - F(x)$.
- What can we expect when we consider different models F estimated using training data D?

Expected error decomposition

- The expected error can be decomposed as

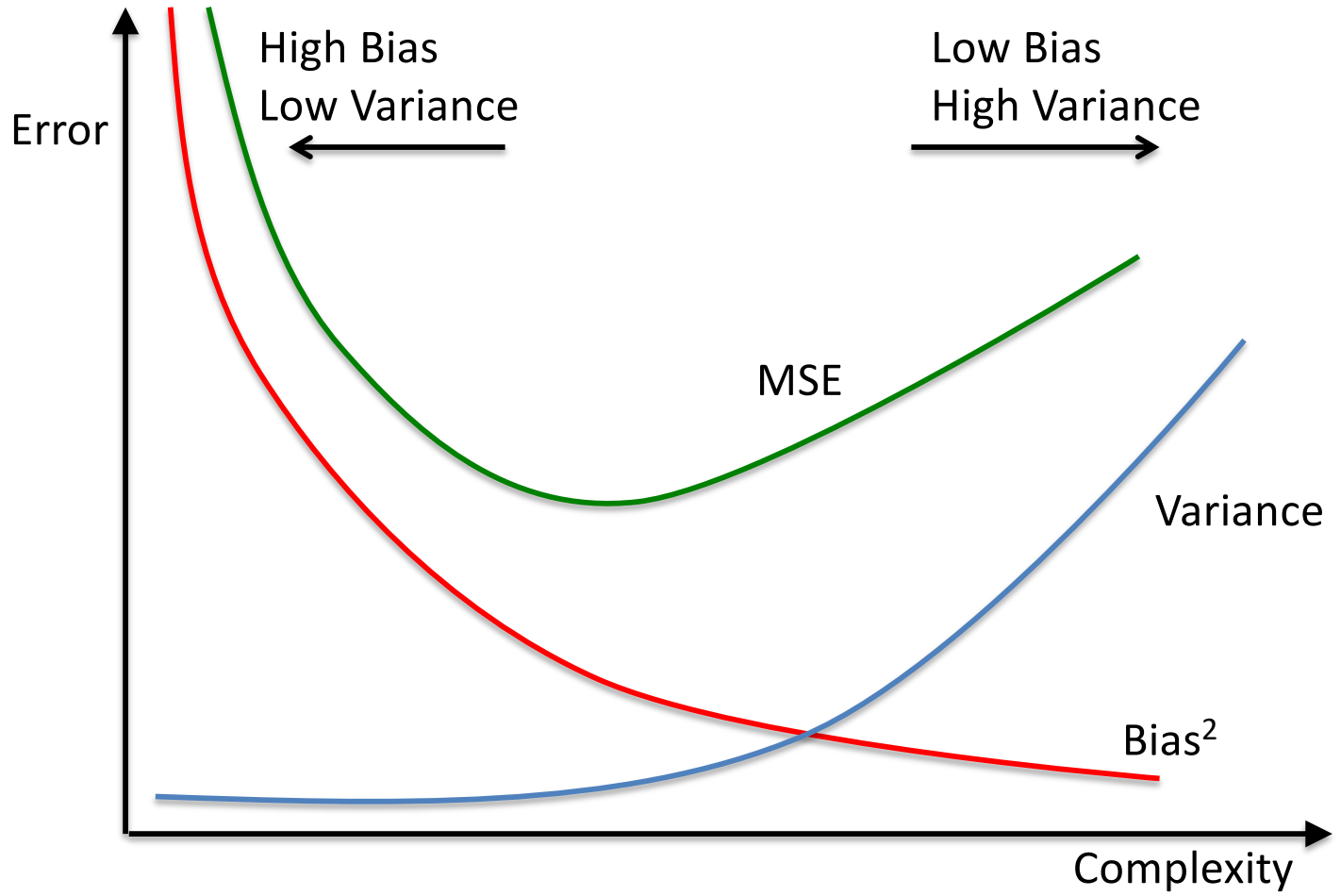
$$E[(y-F(x))^2] = Bias[F(x)]^2 + Var[F(x)] + \sigma^2$$

- Where we have
- $Bias[F(x)] = E[F(x)] - G(x)$
- $Var[F(x)] = E[(F(x)-E[F(x)])^2]$
- σ^2 is known as the irreducible error arising from the noisy measurements

Bias-Variance

- Bias–variance tradeoff (or dilemma) refers to the desire of simultaneously minimizing two sources of error that prevent models from generalizing to new data.
- Bias is error from an inadequate model.
  - High bias is associated with under-fitting: the model fails to detect relevant relations between predictors and responses.
- Variance is error from sensitivity to small fluctuations in the training set.
  - High variance can arise from over-fitting: modeling the noise in the training data, rather than the signal.

Bias-variance



- Out-of-sample MSE reflects both bias and variance.

Source: McSharry (2005), The danger of wishing for chaos

Shrinkage

- Shrinkage relates to the fact that it may be possible to improve a raw estimate by combining it with other information.
- In the case of raw estimates from a particular dataset, the other information may relate to constraints (known *a priori*) on the variable that is being estimated.
- Shrinkage can be used to set some parameters to zero and therefore decrease complexity.

Ridge Regression

- For any fixed non-negative λ, we define the ridge regression criterion as:

$$L(\lambda, \beta) = \left|y - X\beta\right|^2 + \lambda\left|\beta\right|^2$$

- where
- $$\left|\beta\right|^2 = \sum_{j=1}^{p} \beta_j^2$$
- The ridge regression estimator is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\lambda, \beta)$$

Lasso

- For any fixed non-negative λ, we define the Lasso criterion as:

$$L(\lambda, \beta) = \left|y - X\beta\right|^2 + \lambda \left|\beta\right|_1$$

- where
- $$\left|\beta\right|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$$
- The Lasso estimator is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\lambda, \beta)$$

Elastic Net

- For any fixed non-negative λ_1 and λ_2, we define the naive elastic net criterion as:

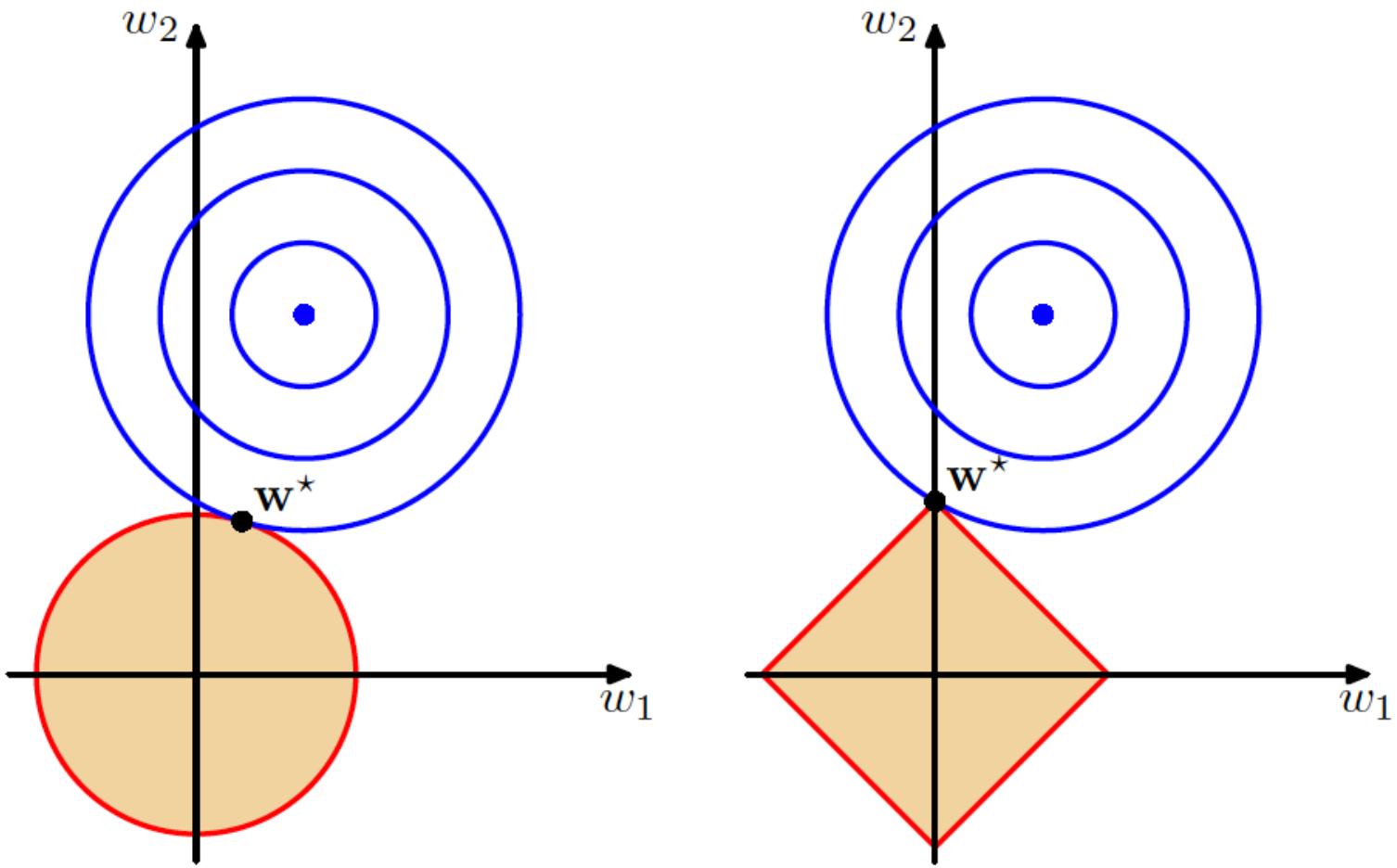
$$L(\lambda_1, \lambda_2, \beta) = \left|y - X\beta\right|^2 + \lambda_2 \left|\beta\right|^2 + \lambda_1 \left|\beta\right|_1$$

- where
- $$\left|\beta\right|^2 = \sum_{j=1}^{p} \beta_j^2, \quad \left|\beta\right|_1 = \sum_{j=1}^{p} \left|\beta_j\right|$$
- The naive elastic net estimator is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\lambda_1, \lambda_2, \beta)$$

Zou & Hastie (2005)

Parameter space



Contours of the unregularized error function (blue) along with the constraint region for the quadratic regularizer q = 2 on the left and the lasso regularizer q = 1 on the right, in which the optimum value for the parameter vector **w** is denoted by **w***.The lasso gives a sparse solution in which $w_1^* = 0$. Source: Chris Bishop, PRML.

Prostate Cancer

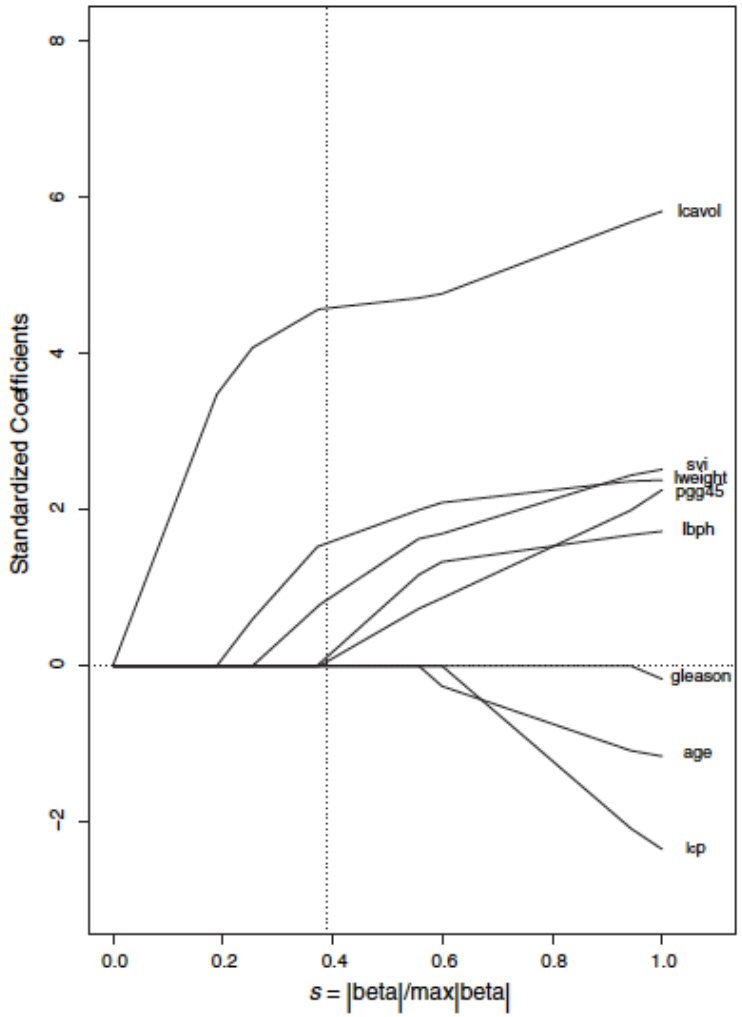
- Prostrate cancer data was originally analysed by Stamey et. al (1989)
- The data was collected from 97 men preparing for prostatectomy
- The objective is to quantify the relationship between cancer volume and eight other clinical measures that are potential candidates as predictors

Prostrate Cancer data

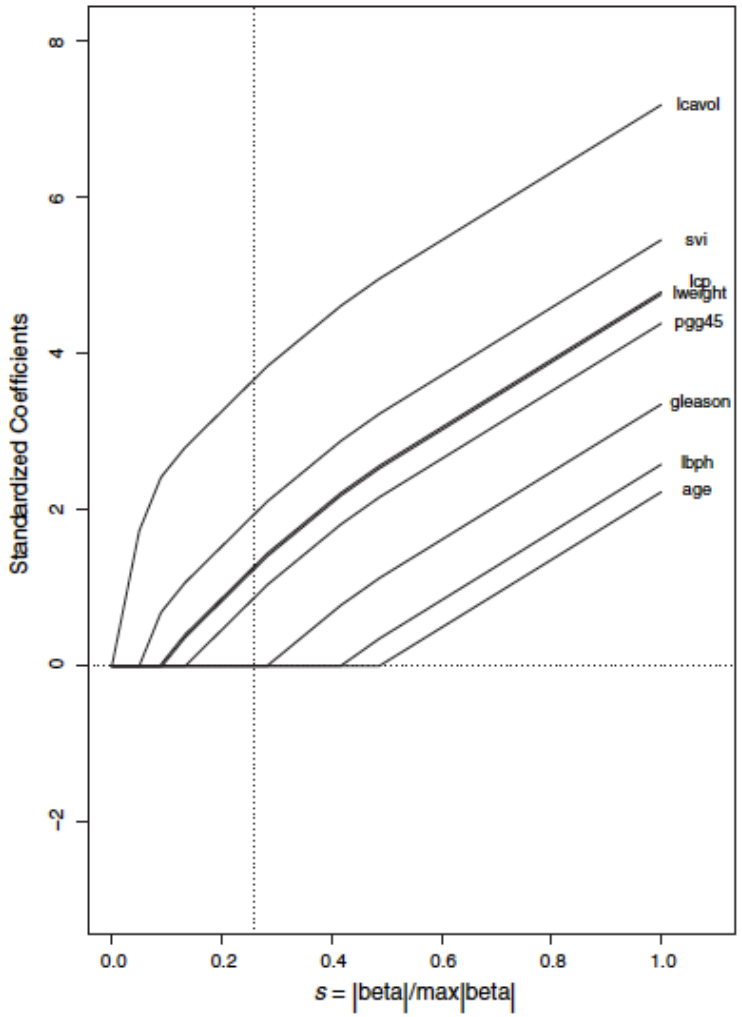
Index	Label	Explanatory variable
1	lcavol	log(cancer volume)
2	lweight	log(prostate weight volume)
3	age	age
4	lbph	log(benign prostatic hyperplasia)
5	svi	seminal vesicle invasion
6	lcp	log(capsular penetration)
7	gleason	Gleason score
8	pgg45	percent Gleason scores 4 or 5
y	lpsa	log(prostate specific antigen)

Prostrate Cancer

Lasso

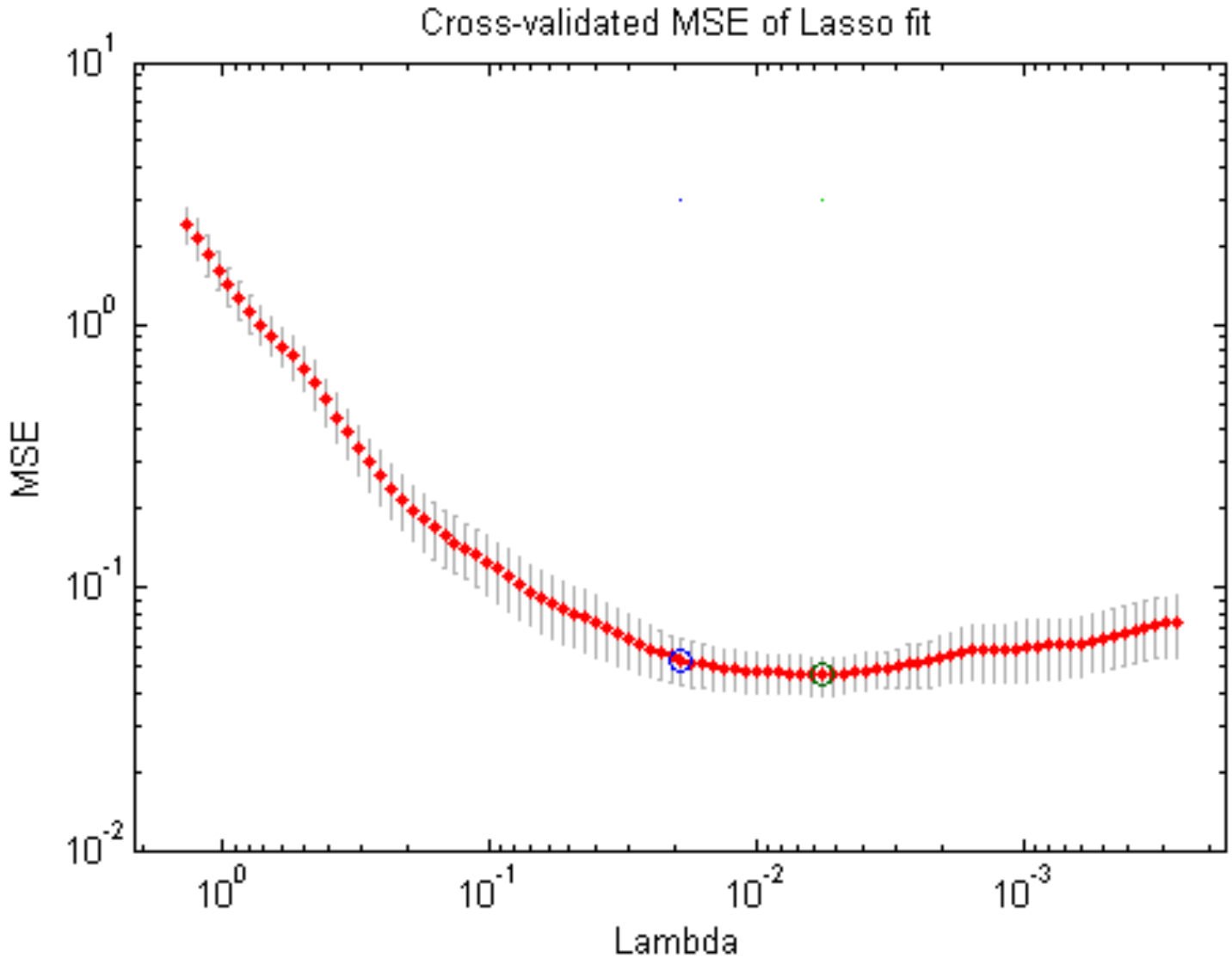


Elastic Net



Zou & Hastie (2005). Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 67, Part 2, pp. 301–320

Optimizing parameters



Source: mathworks

Out-of-sample

- The idea of out-of-sample testing is to avoid the problem of over-fitting when estimating parameters and quantifying performance
- This is achieved by separating the estimation and evaluation procedures:
- **Estimation**: learning data set X_L is used to estimate all model parameters
- **Evaluation**: testing data set X_T is used to calculate performance metrics
- Ensure that there is no overlap between X_L and X_T

Experimental setup

- The simplest approach to out-of-sample testing is when the data is divided into two parts:
- Learning data set for training the model by estimating its parameter values
- Testing data set for evaluating the model and calculating performance metrics
- Disadvantages are that the results depend on the model fit in X_L and the evaluation on X_T

Leave one out

- Suppose there are N observations in the data matrix
- An approach to make better use of the data is to leave out one data point and fit the model to the remaining N-1 observations
- Next a prediction/classification is made for the one observation that was left out
- This procedure is then repeated for all N observations separately
- Finally an average is calculated for the outcomes from all these N out-of-sample evaluations

Cross validation

- The idea of cross validation is to generalize the out-of-sample testing procedure
- Given a dataset with N observations, a partition is made by randomly assigning observations to:
- (1) a learning dataset X_L of size N_L and
- (2) a testing dataset X_T of size N_T
- where $N = N_L + N_T$

k-fold cross validation

- The data is randomly divided into *k* partitions of equal size
- A single partition is selected for testing and the remaining *k*−1 partitions are used as training data for fitting the model
- The cross-validation process is then repeated *k* times (the *k folds*)
- Each of the *k* partitions is used exactly once for evaluation
- The *k* results from the folds are then combined to produce a single summary of performance

Properties of k-fold

- All observations are used for both training and testing
- Each observation is used for testing exactly once
- When *k*=*n* (the number of observations), the *k*-fold cross-validation becomes equal to the leave-one-out cross-validation
- 2-fold cross-validation is also known as the holdout method
- 10-fold cross-validation is commonly used

Repeated random sub-sampling validation

- Randomly splits the dataset into training and testing data
- For each split, the model is fit to the training data, and predictive performance is evaluated using the testing data
- Results are then averaged over the splits
- Advantage: proportion of the training/validation split is not dependent on the number of iterations (folds)
- Disadvantage: overlapping testing subsets imply that some observations may never be selected in the testing subsample, whereas others may be selected more than once

Q&A

Applied Machine Learning

WEEK 8B

Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Constructing models for probabilities	10
2	Discussion	Modelling probability	10
3	Case studies	Heart attack/ Credit risk	10
4	Analysis	Logistic regression	20
5	Demo	Techniques for logistic classification	20
6	Q&A	Questions and feedback	10

Probability

- Probability is a measure of the likeliness that an event will occur, ranging from 0 to 1.
- Probability of 0 implies impossibility and 1 indicates certainty.
- Probability distribution functions (PDFs) convey the probability that any real observation will fall between any two limits.

Quiz

- Based on informaton about COVID (244 million cases and 5 million deaths globally), what is the probability of death inferred from the case fatality rate?

a) 0.5%

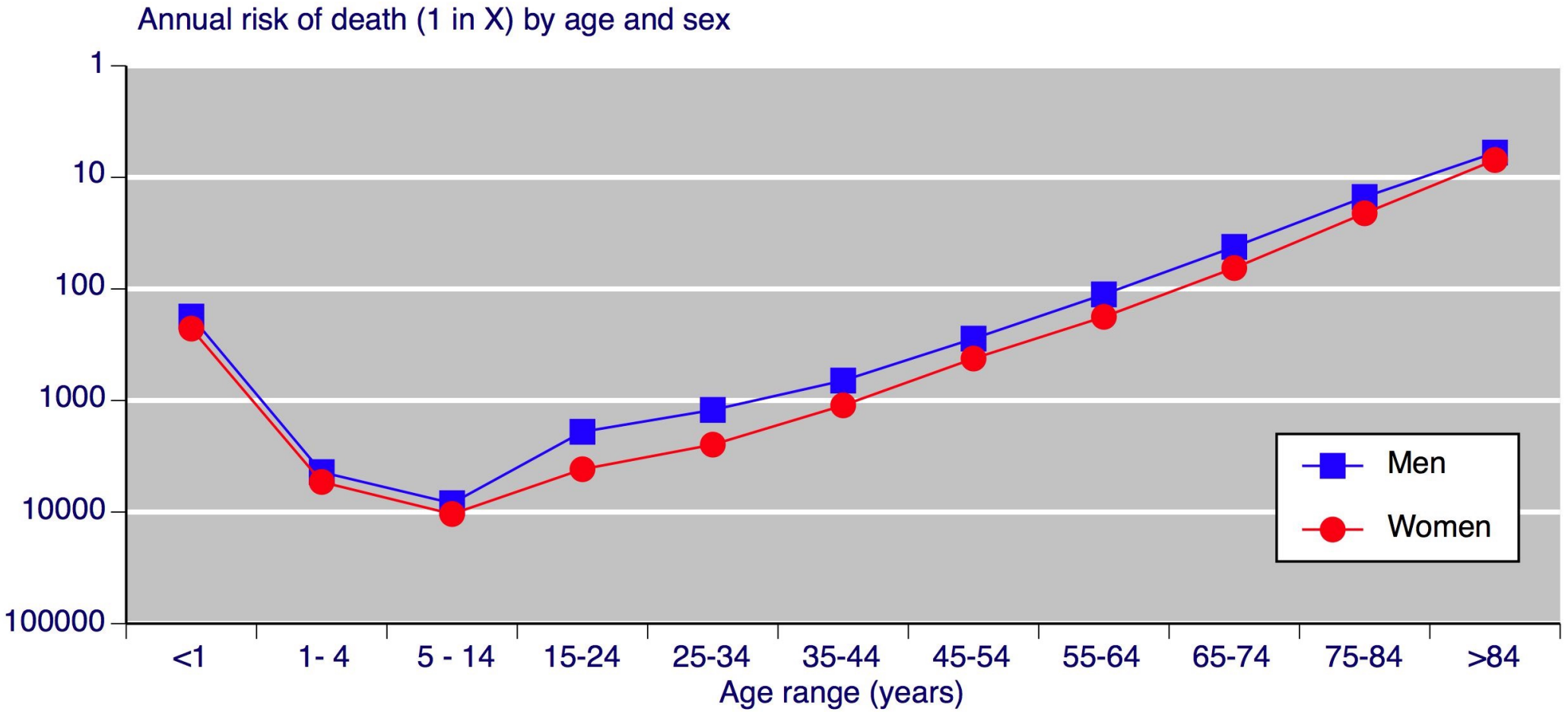
b) 1.0%

c) 2.0%

d) 5.0%

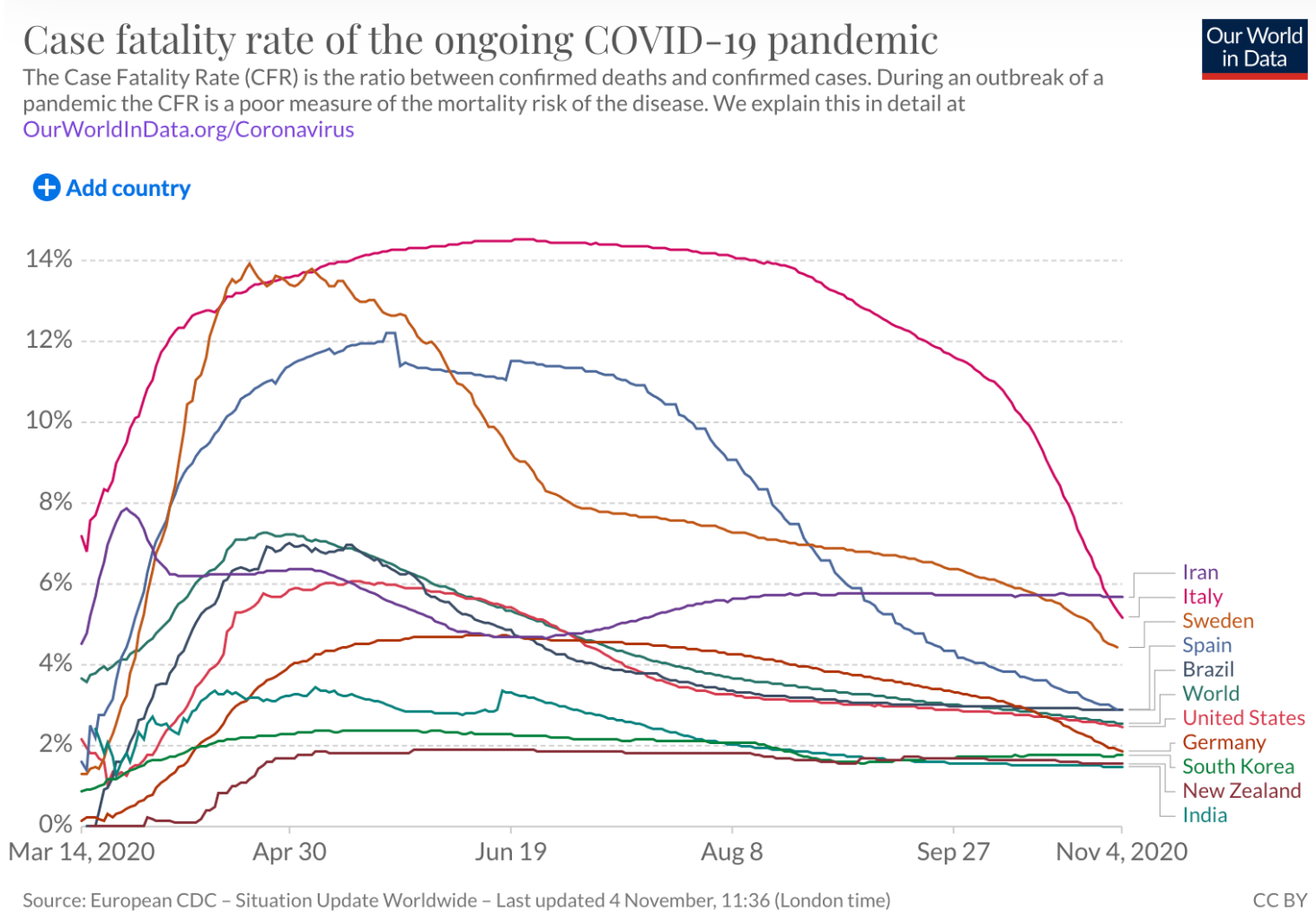
www.Slido.com **#86166**

Annual risk of death by age & sex (UK)



http://www.bandolier.org.uk/booth/Risk/dyingage.html. Note: 2.5% is equivalent to 1 in 40

Case fatality rate for COVID



https://ourworldindata.org/mortality-risk-covid

Bernoulli distribution

- The Bernoulli distribution is the probability distribution of a random variable which takes value 1 with success probability, p, and the value 0 with failure probability, $q = 1-p$.
- If X is a Bernoulli random variable, then
- $P(X=1) = 1 - P(X=0) = 1-q = p$
- Expected value: $E(X) = p$
- Variance: $Var(X) = p(1-p)$

Data: response and predictors

- Consider the response variable:

  $\mathbf{y} = (y_1, \ldots, y_n)^T$

  containing discrete responses.
- For example, in the case of binary classifcation, y could take on values {0,1} for {failure, success}.
- The n x p model matrix **X** defined by:

  $\mathbf{X} = [\mathbf{x}_1 | \ \ldots | \mathbf{x}_p]$

  contains predictors (explanatory variables):

  $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T, \ j = 1, \ldots, p$

Linear classification

- Linear classification refers to the situation where the model structure is linear in the parameters:

$$y = f(X\beta) = f(\Sigma_j \beta_j x_j)$$

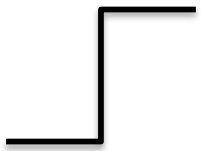
- where $f(.)$ is a function that converts the signal derived from product of the two vectors into the desired output.
- For example, in binary classification, f maps all values above a certain threshold to the first class and all other values to the second class.

Linear models

- Signal from linear relationship: $s = X\beta$
- General form: $y = f(s) = f(X\beta)$
- Three different approaches for producing the final response y depending on application.

Linear classification

$y = \text{sign}(X\beta)$

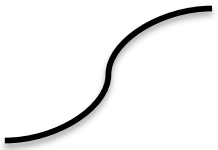


Linear regression

$y = X\beta$

Logistic regression

$y = \text{logistic}(X\beta)$



Poll

- Which variables might predict the probability of a heart attack?

www.Slido.com #86166

Heart attack analysis

- Input variables X could include:

  x_1: age

  x_2: body mass index (weight, height)

  x_3: cholesterol level

  x_4: family history

- $s = X\beta$ gives a risk score
- $y = f(X\beta)$ gives the probability of a heart attack

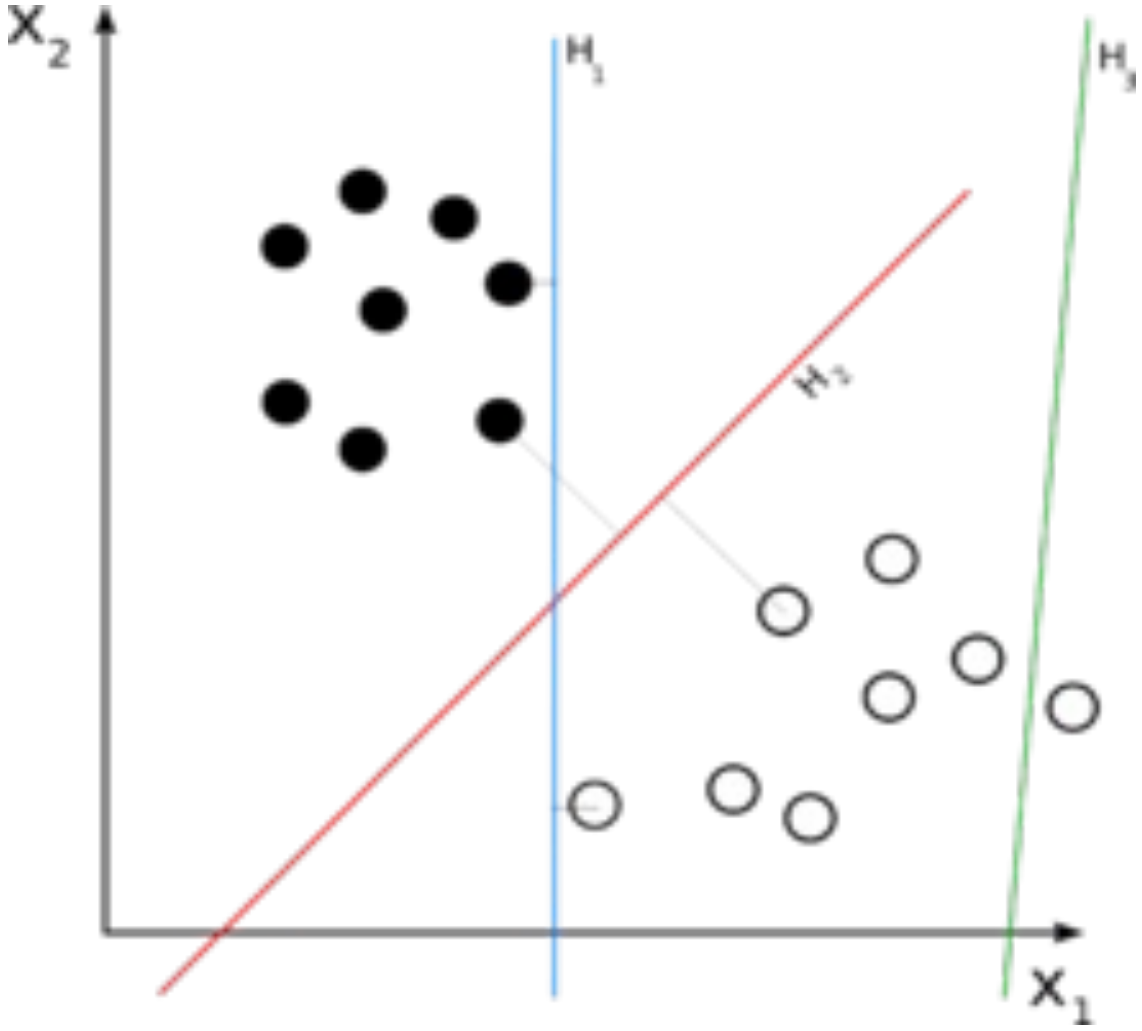
Unit of analysis

Application	Units
Credit Risk	Customers
Credit card fraud	Transactions
Spam detection	Emails
Heart attack	Patients
Football outcome	Games

Credit risk analysis

Approve or Reject?	• Binary classification
How much credit?	• Linear regression
Probability of default	• Logistic regression

Binary classifiers



Poll

- The optimal classifier is:

a) H1
b) H2
c) H3

www.Slido.com **#86166**

Source: wikipedia

Probability of success

- Suppose the probability of success depends on explanatory variable x.
- Constant changes in x are unlikely to cause constant changes in probability.
- This means that the probability of a success is a non-linear function of the explanatory variable x.
- What kind of function would be appropriate?

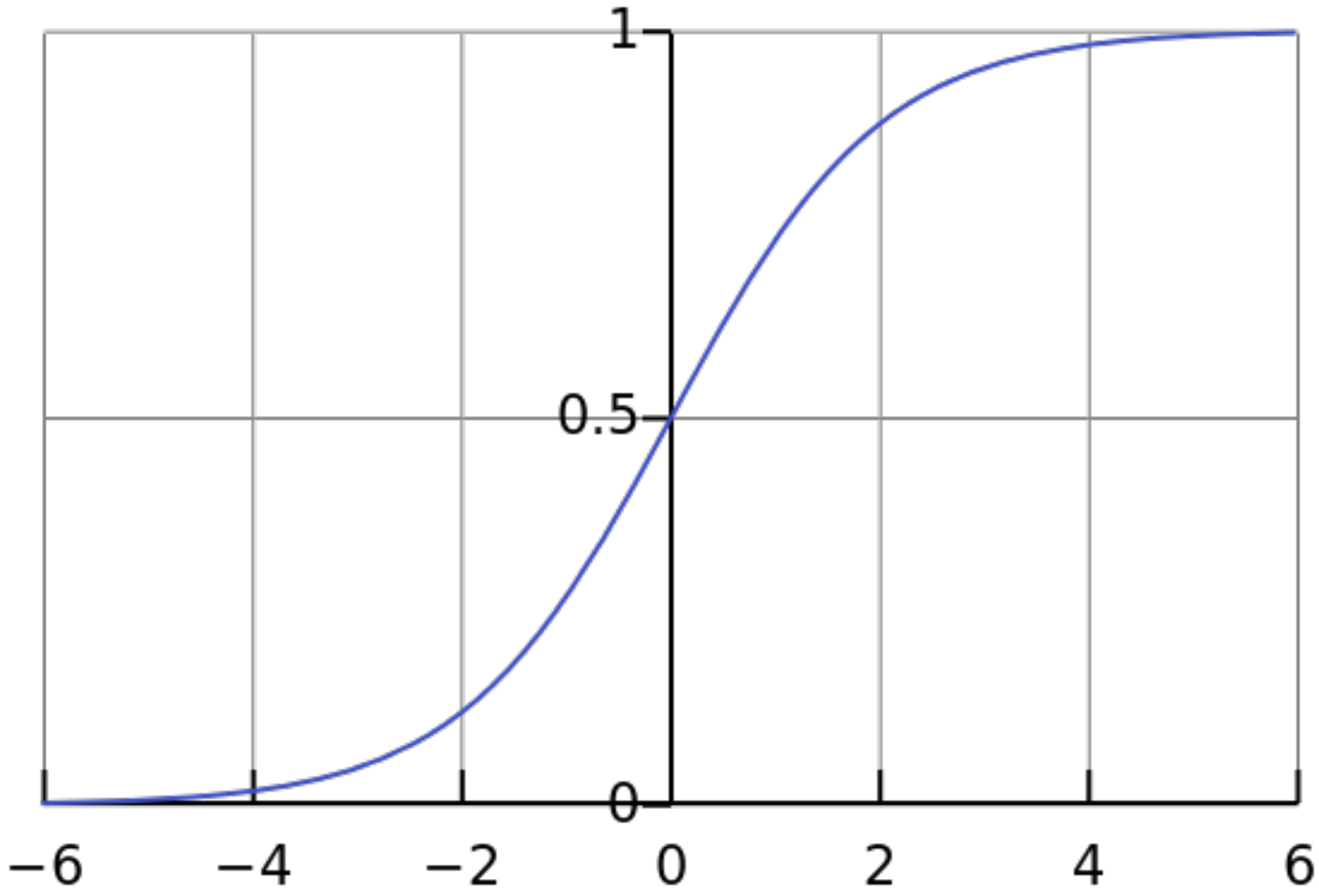
Logistic sigmoid

- The logistic is an S-shaped function which is often referred to as a squashing function.
- It maps the whole real axis [-∞ ∞] into the finite interval [0 1] and is given by

$$y(x) = 1/[1+\exp(-x)]$$

- Alternative: $\tanh(x) = 2y(x)-1$
- Symmetry: $y(-x) = 1- y(x)$
- Inverse: $x = \ln(y/(1-y))$

Logistic function



Logit

- The logit maps the the finite interval [0 1] into the whole real axis [-∞ ∞].
- The logit of a number p between 0 and 1 is given by

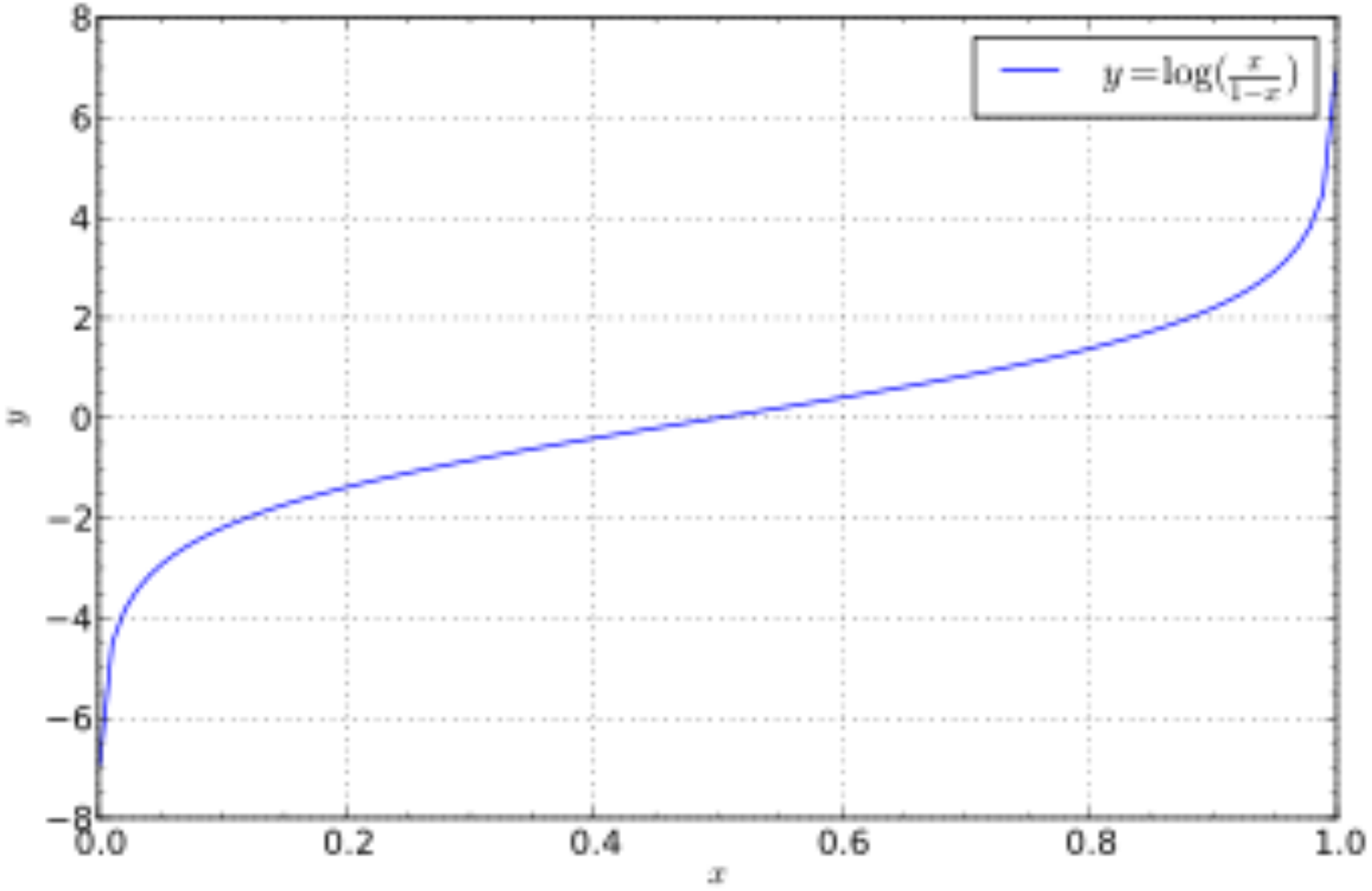
$$\text{logit}(p) = \log(p/(1-p)) = \log(p) - \log(1-p)$$

- Therefore the logistic function of x is the inverse logit given by:

$$\text{logit}^{-1}(x) = 1/[1+\exp(-x)] = \exp(x)/[\exp(x)+1]$$

- The logit in logistic regression is a special case of a link function in a generalized linear model.

Logit function



Logistic regression

- Logistic regression relies on a generalized linear model.
- Logistic regression provides a means of predicting the outcome of a categorical dependent variable based on one or more explanatory variables.
- The dependent variable is binary, indicating the choice between two categories.
- Applications with more than two categories are referred to as multinomial logistic regression.
- Ordered logistic regression refers to cases where the multiple categories are ordered.

Odds

- Odds reflect the likelihood that a particular event will take place or be a success.
- Odds are often expressed as a ratio of the number of outcomes where the event occurs to the number of outcomes where the event does not occur.
- For example, in a game, the odds can be represented as wins:losses denoted by *W:L*.
- Odds in favour are $o_f = W/L$.
- Odds against are $o_a = L/W$.
- Note that $o_f = 1/o_a$, $o_a = 1/o_f$ and $o_a o_f = 1$.

Probabilities from outcomes

- Probability of success, *p*, or failure, *q*, can be computed by dividing by the total number of outcomes.
- Probability of success: $p = W/(W+L) = 1-q$
- Probability of failure: $q = L/(W+L) = 1$-p
- As success and failure are the only possible outcomes, we require that the the probability of success and probability of failure sum to one: $p + q = 1$.

Odds from Probabilities

- Given a probability of success, *p,* the odds as a ratio is expressed as probability of success to probability of failure, which is *p:q* in our notation.
- The odds as numbers can be computed by dividing:
- Odds in favour: $o_f = p/q = p/(1-p) = (1-q)/q$
- Odds against: $o_a = q/p = (1-p)/p = q/(1-q)$

Probabilities from Odds

- If we are given the odds as a number, o_f, then we can derive the probabilities.
- First the odds as a number can be expressed as the odds ratio:

  $o_f:1$ or equivalently $1:(1/o_f) = 1:o_a$

- The probability of success and failure is then:

  $p = o_f/(o_f + 1) = 1/(o_a + 1)$

  $q = o_a/(o_a + 1) = 1/(o_f + 1)$

Odds example

- Consider the days of the week as an example.
- There are two weekend days (Sat and Sun).
- There are five non-weekend days (Mon to Fri).
- There are seven outcomes (all seven days).
- Odds that a randomly chosen day of the week is a weekend are two to five (2:5).
- Odds of a weekend are 2 to 5
- Chances of a weekend are 2 in 7
- Probability of a weekend is $p = 2/7$.

Probability or Odds

- The problem with probabilities is that they are non-linear.
- For example, increasing from 0.1 to 0.2 doubles the likelihood of success, but increasing from 0.8 to 0.9 only slightly increases the likelihood.
- In order to use a linear model, it is more convenient to work with odds.

Odds ratio

- The odds ratio is defined as the ratio of the odds over 1 – the odds.
- This equals the probability of winning over the probability of losing.
- For example, 5 to 1 odds equates to an odds ratio of 0.2/0.8 = 0.25.
- The logit is the natural log of an odds ratio; often called a log odds even though it really is a log odds ratio.
- The logit scale is linear and functions much like a z-score scale.

The logit and probability

- Let p denote the probability of success to be estimated.
- Then the odds ratio for success is p/(1-p).
- The logit is the natural logarithm of the odds ratio and can be expressed as a linear scale:

$$\ln\left[\frac{p}{1-p}\right] = X\beta$$

Probit

- The probit function maps the the finite interval [0 1] into the whole real axis [-∞ ∞].
- The probit function is the quantile function associated with the standard normal distribution.
- The standard normal distribution is commonly written as N(0,1) and its cumulative distribution function as $\Phi(z)$.
- The probit is defined as the inverse of $\Phi(z)$:

$$\text{probit}(p) = \Phi^{-1}(p)$$

Probit regression

- Probit regression uses a probit link function to obtain binary outcomes using a linear combination of the input variables:

$$P(Y=1 \mid X) = \Phi(X\beta)$$

- where Φ is the cumulative distribution function of the standard normal distribution.

Poll

- For a binary classifier, how many distinct outcomes are there when considering a prediction and label?

a) One

b) Two

c) Three

d) Four

www.Slido.com **#86166**

Confusion matrix

Prediction\actual	p	n
p	TP	FP
n	FN	TN

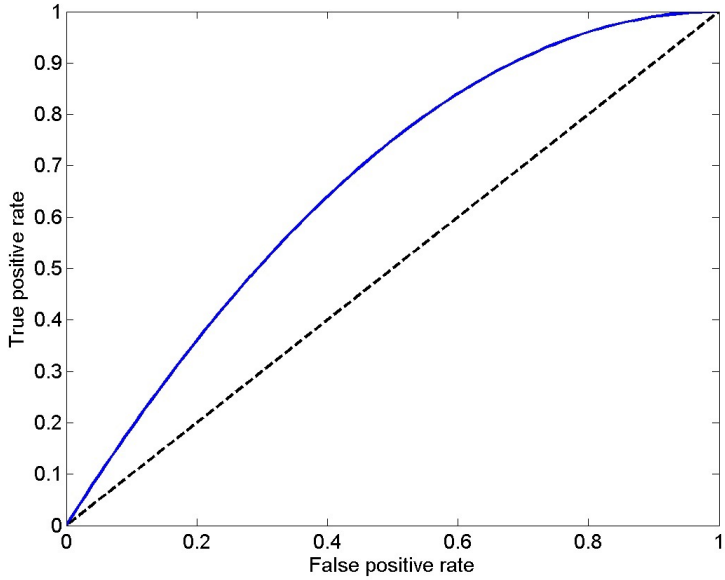
- The two-class prediction problem is also known as binary classification. There are two classes, labeled either as the positive (*p*) or negative (*n*) class
- There are four possible outcomes. If the prediction is *p* and the actual value is also *p*, then it is called a *true positive* (TP); however if the actual value is *n* then it is a *false positive* (FP). In contrast, if the prediction is *n* and the actual value is *n*, it is called a *true negative* (TN); however if the actual value is *n*, it is a *false negative* (FN).
- The four outcomes give a 2×2 *contingency table* or *confusion matrix*

Classification summary

Prediction\actual	p	n
p	TP	FP
n	FN	TN

- True positive rate: TPR = TP / (TP + FN)
- False positive rate: FPR = FP / (FP + TN)
- Specificity =1-FPR
- Sensitivity = Hit Rate = Recall = TPR
- Accuracy: ACC = (TP + TN)/(TP+FN + FP+TN)

ROC analysis



- A receiver operating characteristic (ROC), or simply ROC curve, is a graphical plot of the TPR vs FPR
- The ROC space displays the relative trade-offs between true positive (benefits) and false positive (costs)
- The ROC curve can be summarised using the area under the curve (AUC), which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one
- A good classifier will have an AUC value greater than ½

Multiple categories

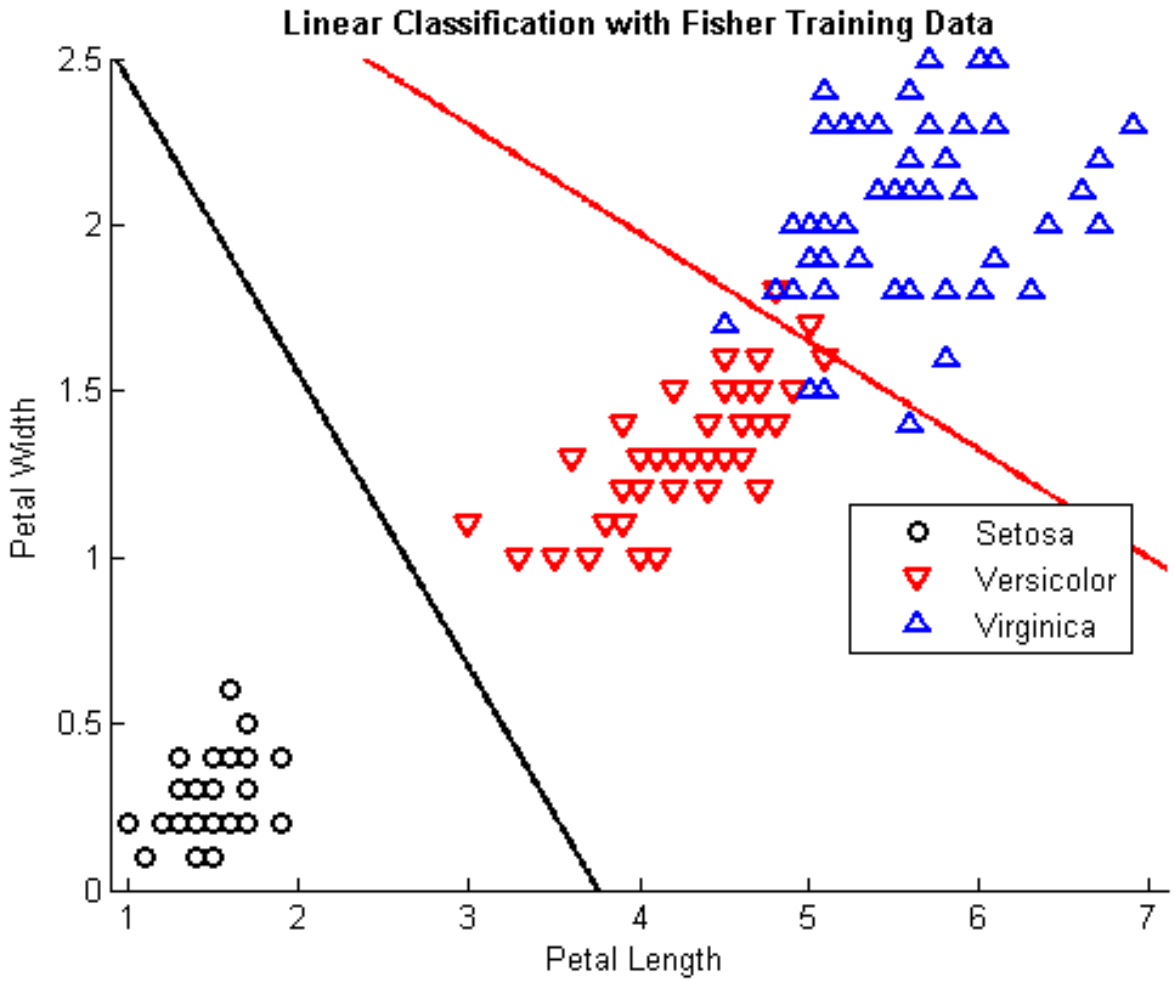
- There are also situations where we wish to distinguish between multiple categories of outcome.
- This is known as multiclass classification.
- For example the objective may be to classify animals based on pictures.
- A multinomial model is required in this case.

Fisher’s Iris data

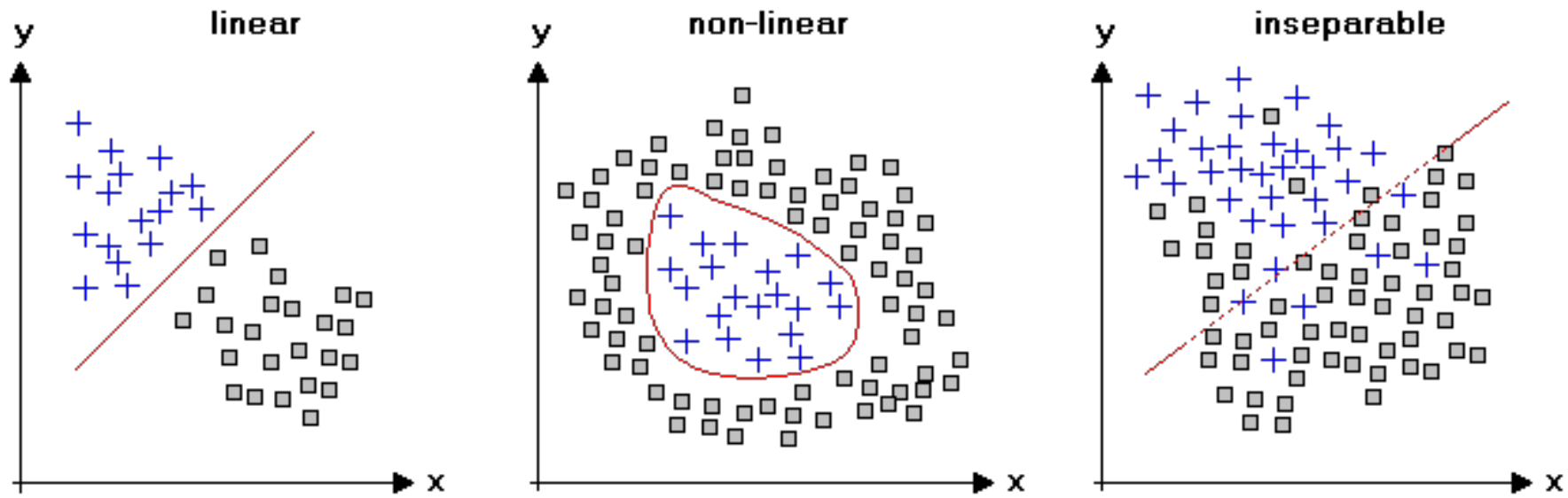
- Ronald Fisher's classification problem consists of three measurements of type
- Type 0 is Setosa; type 1 is Verginica; and type 2 is Versicolor.
- The features are: petal width (PW); petal length (PL); sepal width (SW); and sepal length (SL) for a sample of 150 irises.
- The lengths are measured in millimeters.

Fisher, R. A. (1936). “The Use of Multiple Measurements in Axonomic Problems”, *Annals of Eugenics* **7**, 179-188.

Linear Discriminant Analysis



Classification Challenges



Source: statistics4u.info

Q&A