Data, Inference & Applied Machine Learning

Course: 18-785

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

Fall 2024

ICT Center of Excellence
Carnegie Mellon University

Course outline

Week	Description
1	Regression
2	Feature selection
3	Nonlinear techniques
4	Supervised learning
5	Unsupervised learning
6	Ensemble approaches

Applied Machine Learning

WEEK 10A

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Prediction and classification	10
2	Discussion	Bayes Theorem	10
3	Case study	Spam detection	10
4	Analysis	CART	20
5	Demo	Constructing and evaluating trees	20
6	Q&A	Questions and feedback	10

Poll 1

- Supervised learning is not typically used for:

a) Predicting probability of winning a game
b) Segmenting customers
c) Forecasting energy demand
d) Credit scoring in microfinance

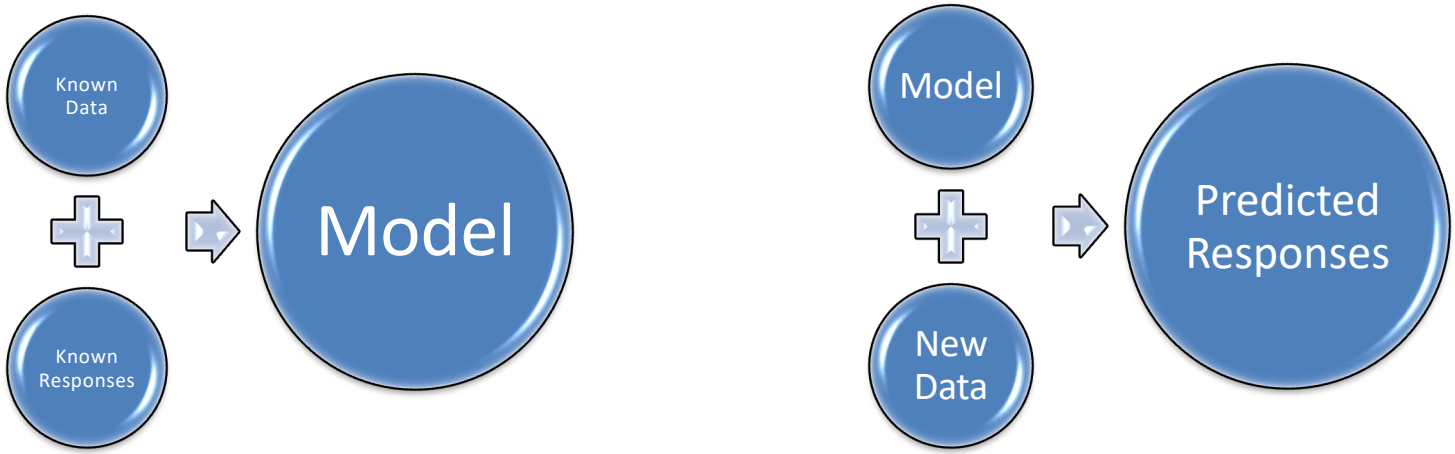
Slido.com
#69159

Supervised Learning

- Supervised learning splits into two broad categories:
- Classification for responses that can have just a few known values, such as 'true' or 'false'. Classification algorithms apply to nominal, not ordinal response values.
- Regression for responses that are a real number, such as miles per gallon for a particular car.

Supervised Learning

- Supervised learning takes a known set of input data and known responses to the data, and seeks to build a predictor model that generates reasonable predictions for the response to new data



Supervised Learning

- Discriminant Analysis
- Naïve Bayes
- Decision Trees
- Neural Networks
- Radial Basis Function
- Kernel methods
- K Nearest Neighbour
- Support Vector Machines

Classification

- One approach to classification is to construct a discriminant function that directly assigns each vector **x** to a specific class.
- An alternative approach is to model the conditional probability distribution $p(C_k|\mathbf{x})$ in an inference stage, and then subsequently use this distribution to make optimal decisions.

Discriminant analysis

- A discriminant is a function that takes an input vector **x** and assigns it to one of *K* classes, denoted C_k.
- In the case of linear discriminants, the decision surfaces that provide classifications are hyperplanes.
- For binary classification, one hyperplane is required.

Bayes theorem - probabilities

- Suppose that p(x) and p(y) are the probabilities of x and y independent of each other.
- Conditional probability p(x|y) is the probability of x given that y is true.
- Conditional probability p(y|x), is the probability of y given that x is true.
- These probabilities are related via:

$$p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x)$$

- Bayes theorem states that

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

Bayes Theorem - Statistical Inference

- Bayes' theorem relates current probability to prior probability.
- For proposition *y* and evidence *x*:
- The *prior* $p(y)$ is the initial degree of belief in *y*.
- The *posterior* $p(y|x)$ is the degree of belief having accounted for *x*.
- The quotient $p(x|y)/p(x)$ represents the support *x* provides for *y*.

Picnic planning

- You are planning a picnic today, but the morning is cloudy, what is the probability of a rained off picnic today?
- Information:
  - 50% of all rainy days start off cloudy: p(cloud|rain) = 0.5
  - But cloudy mornings are common with about 40% of days start cloudy: p(cloud) = 0.4
  - And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%): p(rain) = 0.1
- Bayes formula helps us to take account of all this information and states:

  P(rain|cloud) = p(rain) p(cloud|rain)/p(cloud)

  P(rain|cloud) = 0.1*0.5/0.4 = 0.125
- Recommended decision is to move ahead with picnic plan as 12.5% is a relatively small risk worth taking.

https://www.mathsisfun.com/data/bayes-theorem.html

Poll 2

- You have been in close contact with someone with confirmed COVID in the last week but fortunately a PCR test is negative, your actual probability of COVID infection could be more than 70%.

- True or False

- **Slido.com**
- **#69159**

Interpreting COVID-19 Test Results

- Polymerase chain reaction (PCR) assays test from nasal and pharyngeal swabs has specificity = 99.9% and sensitivity of 70%.
- Specificity is the percentage of negatives correctly identified and sensitivity is the percentage of positives that are correctly identified.

Patient	Pre-test Prob (%)	PCR Sensitivity (%)	Post-test Prob(COVID) after + test result	Post-test Prob(COVID) after – test result
High pre-test probability	90	70	100	73.0
Low pre-test probability	10	70	98.7	3.2

- When COVID-19 infection is likely, such as in a healthcare worker with significant exposure, a negative test should not rule out acute infection. In this case, as recommended by the CDC, repeat testing or further evaluation should be considered.

Good et al. (2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7269418/
https://medium.com/cantors-paradise/bayes-theorem-and-corona-virus-testing-43e2163c925e

Bayes Classification

- Consider spam classification where the two classes are spam and ham.
- We can estimate p(y) by counting the number of spam and ham emails:

  $p(spam) = n_{spam}/n$

  $p(ham) = n_{ham}/n$
- Without knowing p(x), we can form a likelihood:

$$L(x) = \frac{p(spam \mid x)}{p(ham \mid x)} = \frac{p(x \mid spam)p(spam)}{p(x \mid ham)p(ham)}$$

- and declare an email is spam when L(x) exceeds a threshold.

Spam detection

- Consider a filter where each email is labeled as either spam or not spam (ham).
- Suppose that 1% of all emails are spam.
- The imperfect filter outputs positive or negative labels for each email.
- The filter indicates a correct positive result in 98% of emails which are actually spam.
- The filter indicates a correct negative result in 97% of emails which are not spam.

Calculating probabilities

- We know that:
- p(spam) = 0.01 p(ham) = 0.99
- p(pos|spam) = 0.98 p(neg|spam) = 0.02
- p(neg|ham) = 0.97 p(pos|ham) = 0.03
- What should we conclude about a new email which is labeled as spam?
- p(pos|spam)p(spam) = 0.98*0.01 = 0.0098
- p(pos|ham)p(ham) = 0.03*0.99 = 0.0297

Probability of spam

- Note that there are only two outcomes such that the email is either spam or ham.
- p(spam|pos) + p(ham|pos) = 1
- Therefore we have:
- p(spam|pos) = 0.0098/(0.0098+0.0297)=0.2481
- p(ham|pos) = 0.0297/(0.0098+0.0297)=0.7519
- Without knowing p(pos), we have been able to reach the conclusion that this email is most likely not spam.

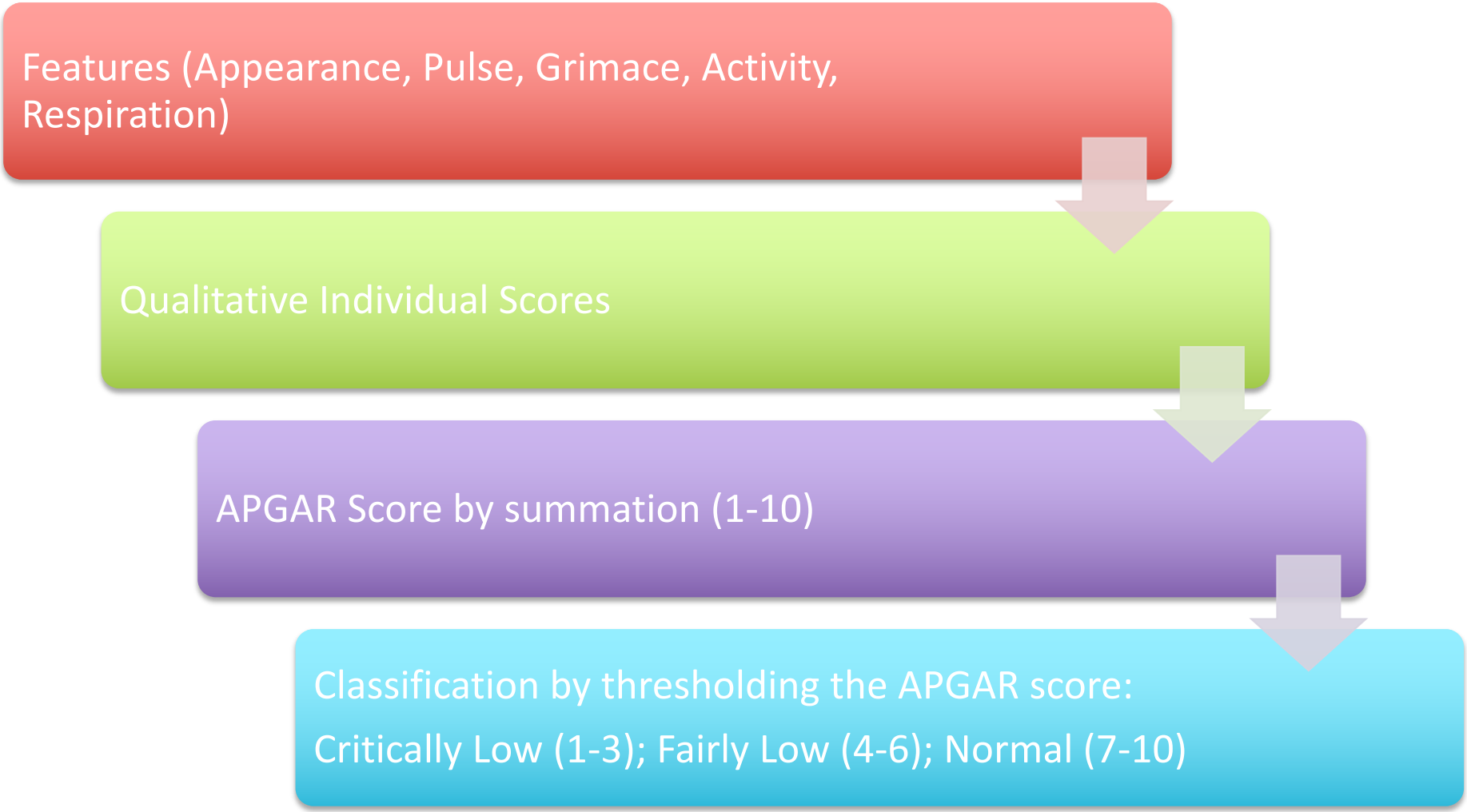
Naïve Bayes

- Consider the construction of a spam classifier using the words in an email as features.
- We estimate p(x|y), that is p(x|spam) and p(x|ham).
- The naïve Bayes approach treats feature values as independent given the outcome.
- As an email x consists of N words w_n, we can write:
- $$p(x \mid y) = \prod_{n=1}^{N} p(w_n \mid y)$$
- p(w|y) can be obtained by counting the frequency occurrence of the word w within emails which are spam and ham.

Five criteria of Apgar score

Component\Score	Score of 0	Score of 1	Score of 2
Appearance	Blue or pale all over	Blue at extremities Body pink	Body and extremities pink
Pulse	Absent	<100	>100
Grimace	No response to stimulation	Grimace on suction or aggressive stimulation	Cry on stimulation
Activity	None	Some flexion	Flexed arms and legs that resist extension
Respiration	Absent	Weak, irregular, gasping	Strong, lusty, cry

APGAR decisions



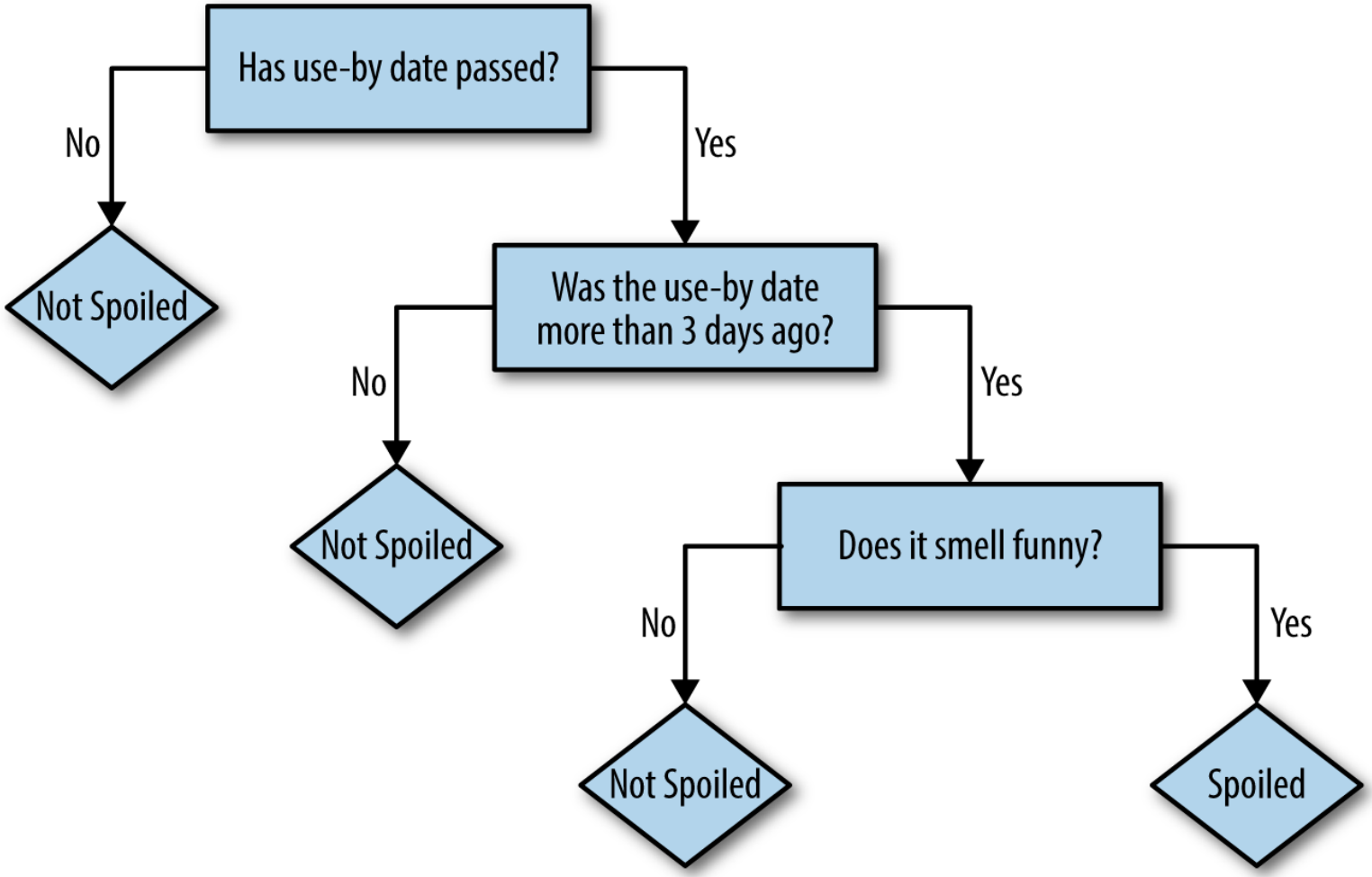
Decision Tree

- A decision tree provides a simple means of making decisions based on the observed values of a number of input features.
- The resulting model is usually nonlinear in structure because of the partitioning of the input feature space.
- Despite the complexity of the model, a decision tree can be relatively easy to use in practice.

Using a Decision Tree

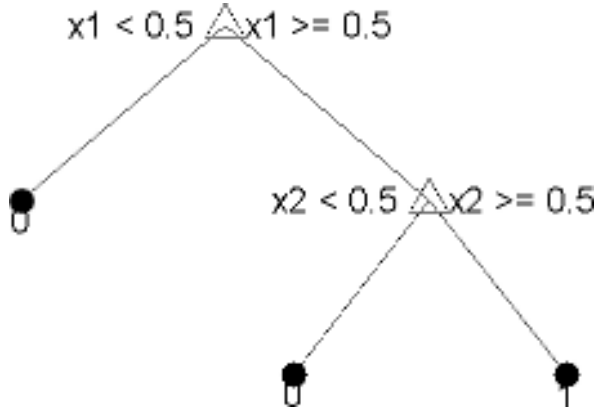
- Classification trees and regression trees predict responses to input features.
- To determine a response, follow the decisions in the tree from the root (beginning) node down to a leaf node.
- The leaf node contains the response.
- Classification trees give responses that are nominal, such as 'true' or 'false'.
- Regression trees give numeric responses.

Deciding what to eat: Is it spoiled?



Source: Josh Wills, Sean Owen, Uri Laserson, Sandy Ryza, “Advanced Analytics with Spark”.

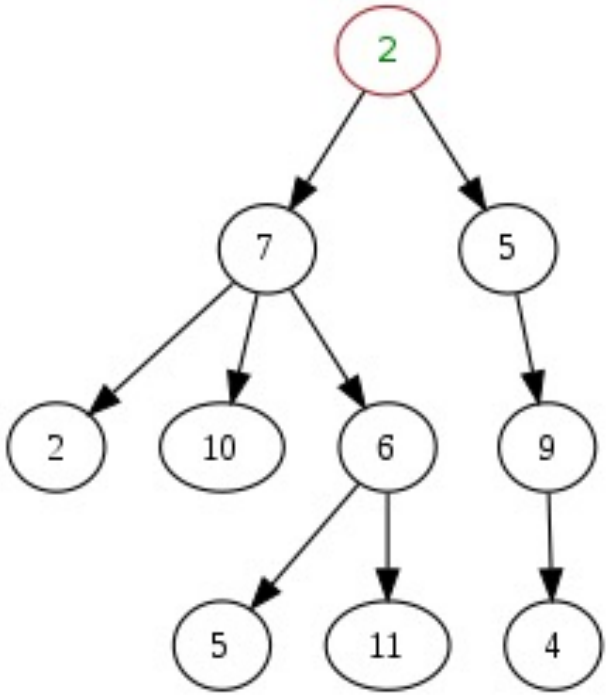
Predictions



- Each step in a prediction involves checking the value of one predictor .
- This tree predicts classifications based on two predictors, x1 and x2.
- To predict, start at the top node, represented by a triangle (Δ).
- The first decision is whether x1 is smaller than 0.5.
- If so, follow the left branch, and see that the tree classifies the data as type 0.
- If, however, x1 exceeds 0.5, then follow the right branch to the lower-right triangle node.
- Here the tree asks if x2 is smaller than 0.5.
- If so, then follow the left branch to see that the tree classifies the data as type 0.
- If not, then follow the right branch to see that the that the tree classifies the data as type 1.

Tree

- A tree is a widely used abstract data type that simulates a hierarchical tree structure.
- A tree has a root value and subtrees of children with a parent node, represented as a set of linked nodes.
- In this diagram, the node labeled 7 has three children, labeled 2, 10 and 6, and one parent, labeled 2. The root node, at the top, has no parent.



Tree Construction

- Decision trees are constructed using the following steps:
- Start with all input data and examine all possible binary splits on every predictor.
- Select a split with best optimization criterion.

  e.g. If the split leads to a child node having too few observations (less than the MinLeafSize parameter), select a split with the best optimization criterion subject to the MinLeafSize constraint.
- Impose the split & repeat recursively for the two child nodes.
- MaxNumSplits — The maximal number of branch node splits is MaxNumSplits per tree. Set a large value for MaxNumSplits to get a deep tree. Default is N-1.
- MinLeafSize — Each leaf has at least MinLeafSize observations. Set small values of MinLeafSize to get deep trees. The default is 1.
- MinParentSize — Each branch node in the tree has at least MinParentSize observations. Set small values of MinParentSize to get deep trees. The default is 10.

Stopping Rule

- Stop splitting when any of the following hold:
- The node is pure.
- For classification, a node is pure if it contains only observations of one class.
- For regression, a node is pure if the mean squared error (MSE) for the observed response in this node drops below the MSE for the observed response in the entire data multiplied by the tolerance on quadratic error per node (qetoler parameter).
- There are fewer than MinParentSize observations in this node.
- Any split imposed on this node would produce children with fewer than MinLeafSize observations.

Optimization Criteria

- Regression: mean-squared error (MSE). Choose a split to minimize the MSE of predictions compared to the training data.

- Classification: the most common is the Gini index. The Gini index is maximum when all observations are equally distributed across the classification classes and minimum when all observations belong to the same class.

Poll 3

- Is overfitting likely a problem for constructing decision trees?

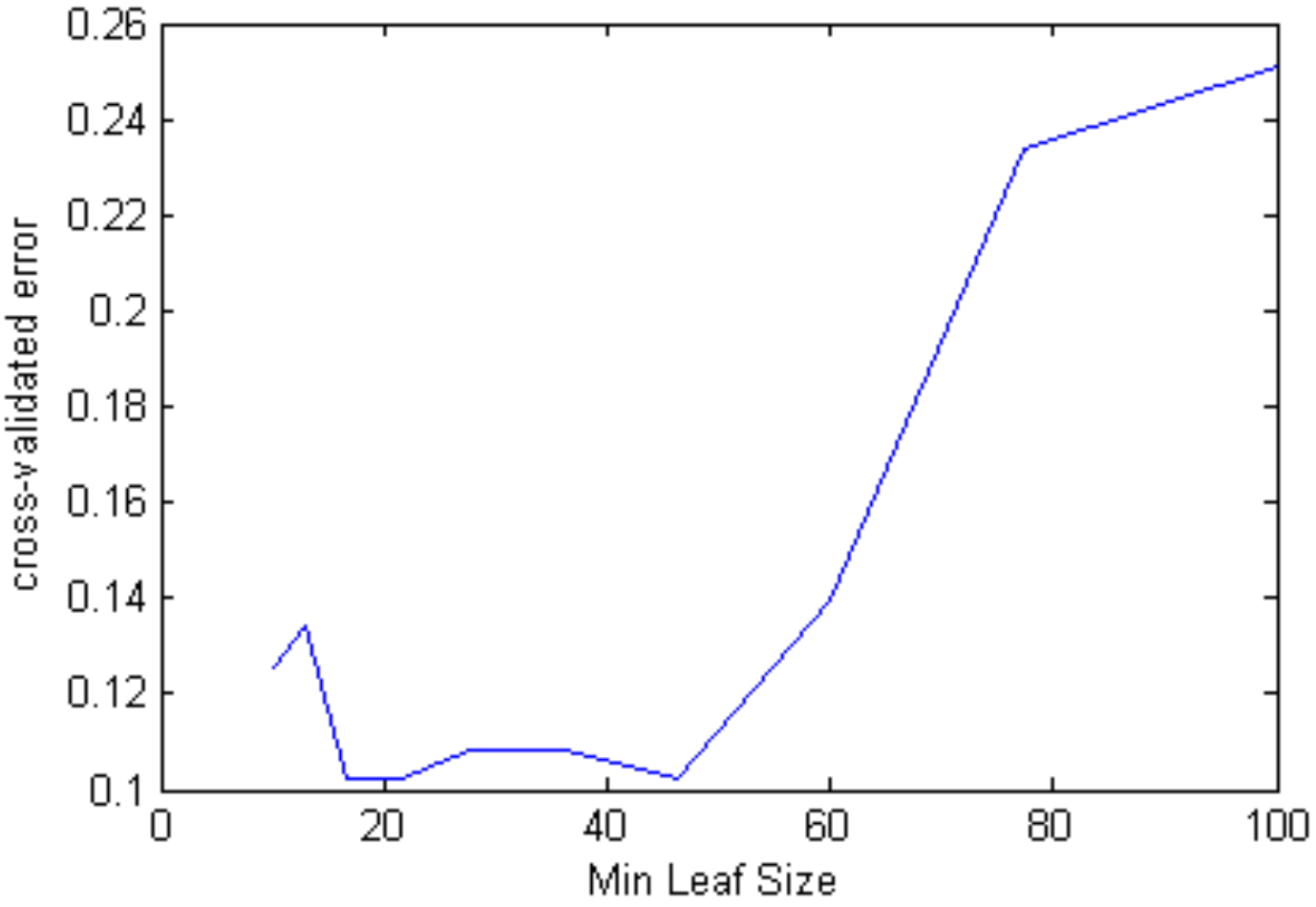
- Yes
- No

- **Slido.com**
- **#69159**

Tree depth or leafiness

- When growing a decision tree, consider its simplicity and predictive power.
- A deep tree with many leaves is usually highly accurate on the training data.
- However, the tree is not guaranteed to show a comparable accuracy on an independent test set.
- A leafy tree tends to overfit, and its test accuracy is often far less than its training accuracy.
- In contrast, a shallow tree does not attain high training accuracy.
- But a shallow tree can be more robust — its training accuracy could be close to that of a representative test set.
- Also, a shallow tree is easy to interpret.

Selecting Appropriate Tree Depth



The final tree

- The decision tree is constructed so that it predominantly uses variables that are relevant for determining the dependant variable.
- The first decision nodes highlight the variables that are most important in this non-linear context.
- Optimal pruning eventually prunes away variables that have little or no information about the dependent variable.

Human neural network

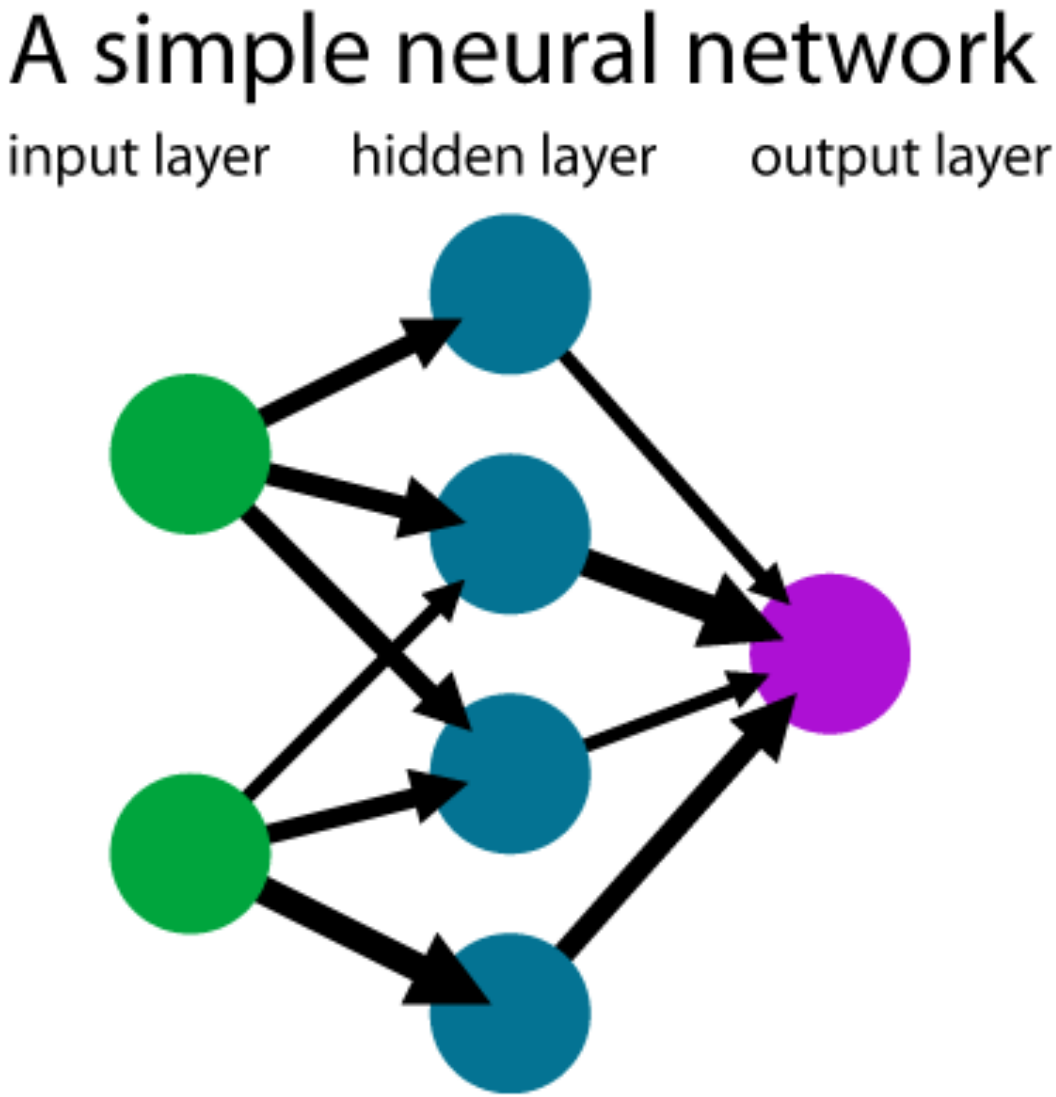
Source: www.extremetech.com

Perceptron

- A perceptron takes real valued inputs, calculates a linear combination, and outputs a 1 if the result is higher than a threshold and -1 otherwise:
- Defining x_0=1 gives

$$y(x_1,\ldots,x_n) = sign(w_0 + w_1 x_1 + \cdots w_n x_n) = sign\left(\sum_{i=0}^{n} w_i x_i\right)$$

- Training a perceptron therefore involves estimating the parameters $w_o,\ldots,w_n$.



Source: wikipedia

Radial basis functions

- Radial basis functions (RBFs) can be viewed as a single-layer neural network
- Select K centres in the state space $\boldsymbol{\xi}_j$ for $j = 1, \ldots, K$
- Choose basis function $\Phi(r)$, examples are: r^3, $\exp(-r^2/2\sigma^2)$
- Model is expressed as a linear weighted sum of the RBFs:

$$F(s_i) = \sum_{j=1}^{K} a_j \phi\left(\left\| s_i - \xi_j \right\|\right)$$

- Model parameters $\boldsymbol{a}$ are determined by solving the linear system of equations $\boldsymbol{b} = \boldsymbol{Ha}$, where the elements of the design matrix are $H_{ji} = \Phi(\|\boldsymbol{s}_i - \boldsymbol{\xi}_j\|)$, and the elements of the dependent variable are $b_i = s_{i+1}$

Neural networks

- Feed-forward network with one hidden layer
- Comprises one layer of input units, one layer of neurons and one layer of output units:

$$F(\mathbf{s}_i) = \sum_{j=1}^{K} \frac{a_j}{1 + \exp\left(\mathbf{b}_j . \mathbf{s} - c_j\right)}$$

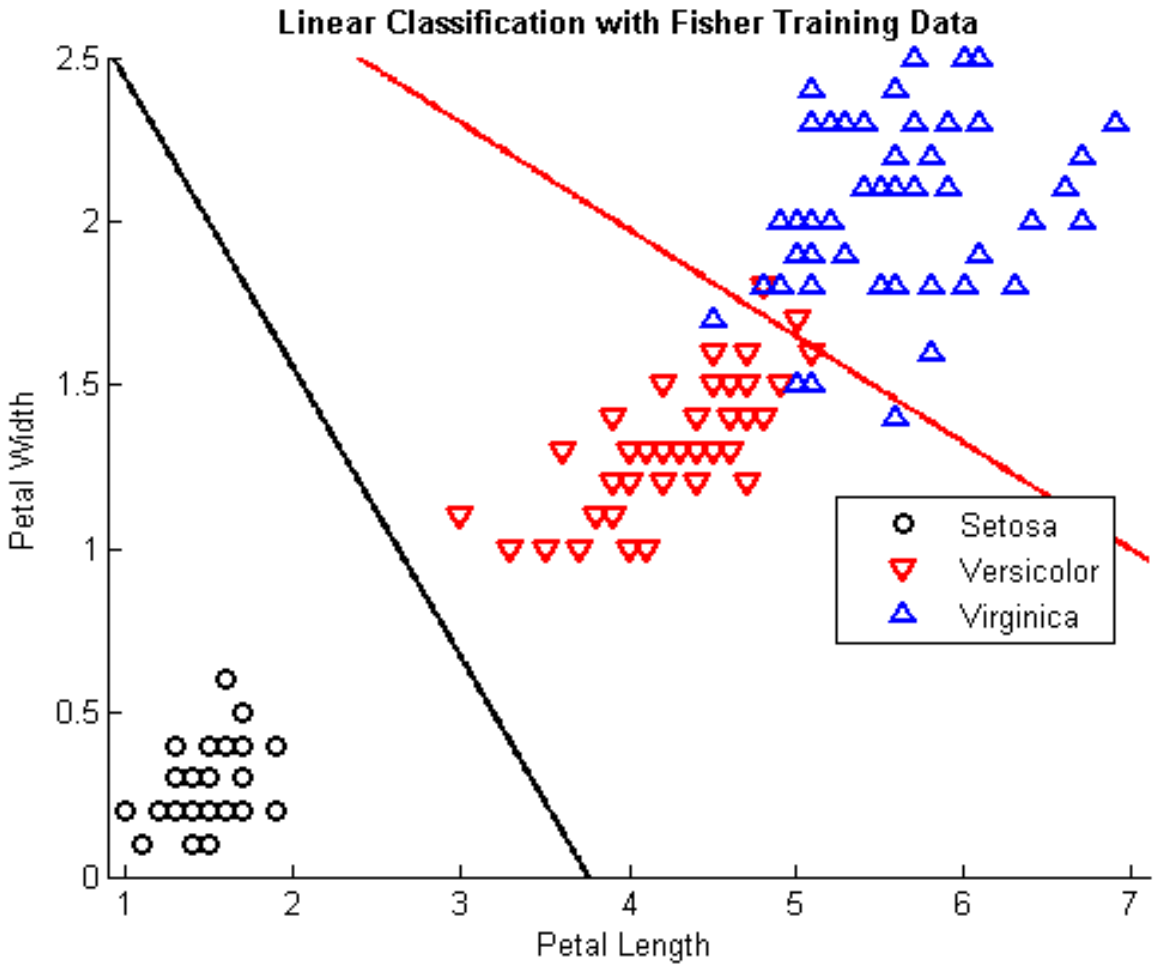
- Where parameters a_i, $\boldsymbol{b}_i$, c_i are determined by nonlinear minimisation techniques
- This NN is simply a global nonlinear function with basis
  $\Phi(\boldsymbol{s}) = 1/(1 + exp(\boldsymbol{b}\cdot\boldsymbol{s} - c))$
- Advantages:
  - Extremely flexible, universal approximators
  - Many easy to use off-the shelf packages
- Disadvantages:
  - Large computational effort for estimating parameters
  - Problems with local minima in parameter space
  - The large number of parameters may lead to over-fitting

Illustration using Fisher's data

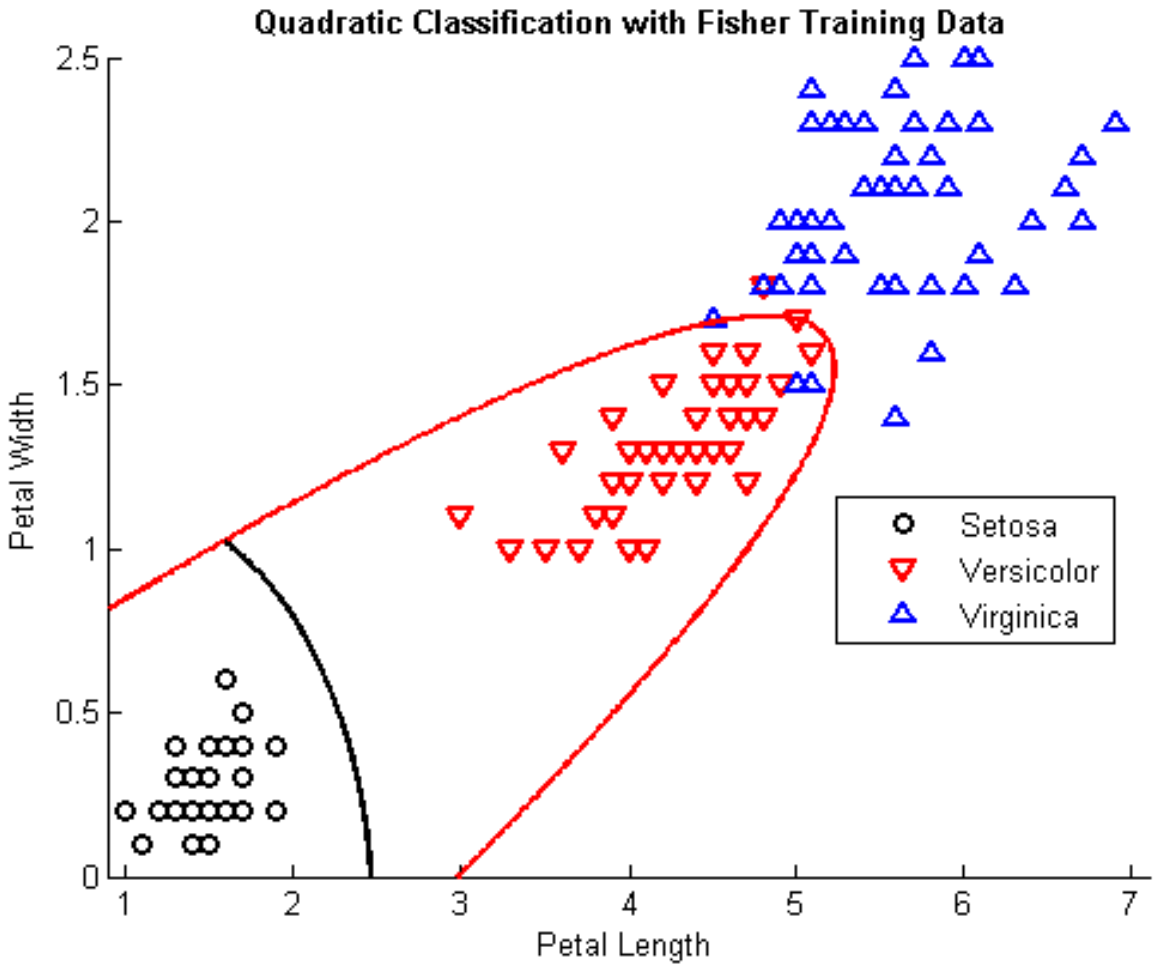
- Ronald Fisher's classification problem consists of three measurements of type
- Type 0 is Setosa; type 1 is Verginica; and type 2 is Versicolor.
- The features are: petal width (PW); petal length (PL); sepal width (SW); and sepal length (SL) for a sample of 150 irises.
- The lengths are measured in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

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

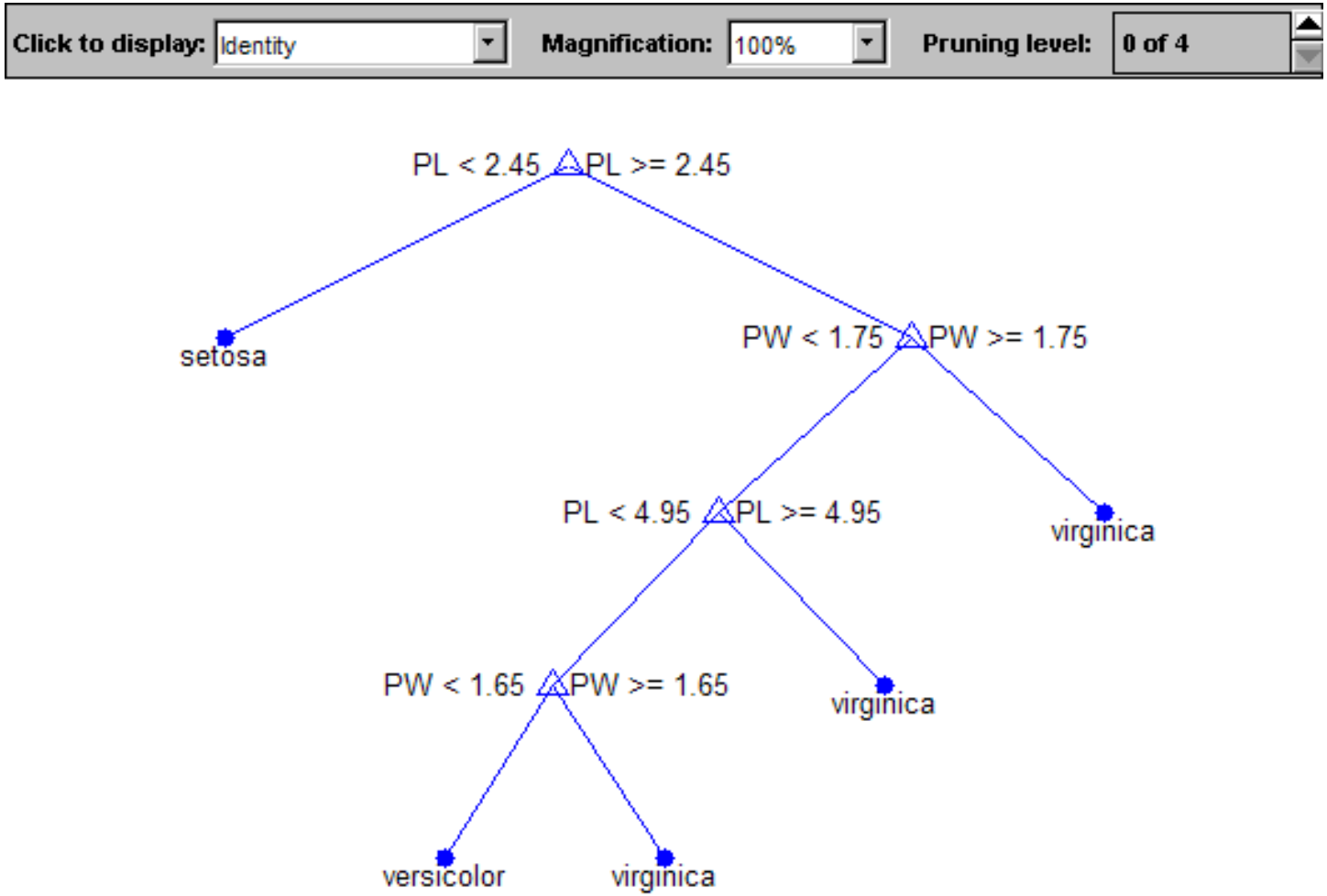
Linear Discriminant Analysis



Quadratic Discriminant Analysis



Example of a Decision Tree



A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences.

Q&A

Applied Machine Learning

WEEK 10B

Today's Lecture

No.	Activity	Description	Time
1	Challenge	Data-driven model structure	10
2	Discussion	Density estimation	10
3	Case study	Bimodal distribution	10
4	Analysis	Non-parametric models	20
5	Demo	KNN	20
6	Q&A	Questions and feedback	10

Summarizing data

- Calculating the average indicates the expected value of a distribution.
- The standard deviation provides information about the spread of the distribution.
- In order to communicate the range of likely outcomes, we might want to visualize the entire distribution.
- What is the best way to model an entire distribution?

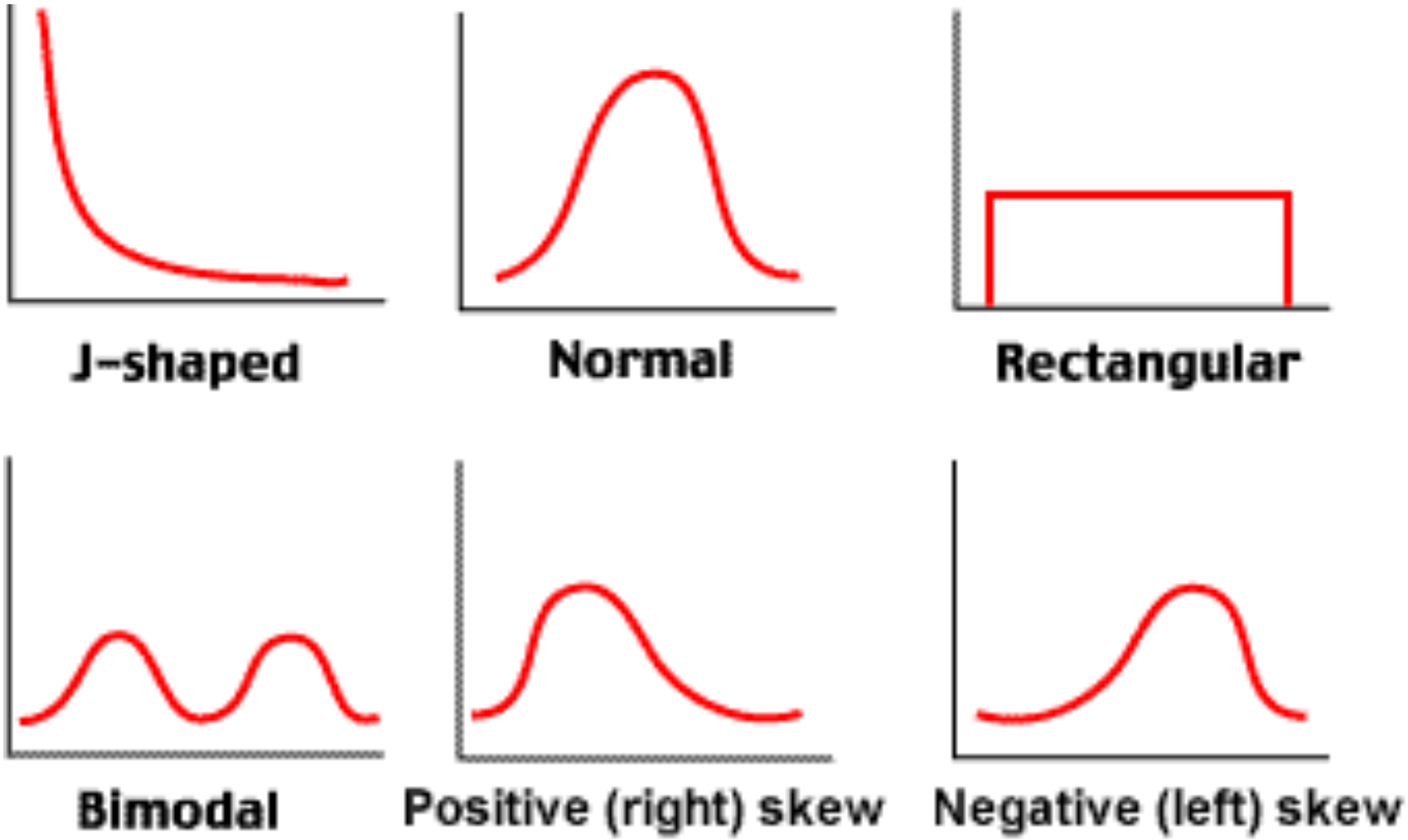
Parametric methods

- Parametric methods rely on a particular functional form (mathematical equation) and a few parameters to describe the data.
- For example, fitting a normal distribution using a series of observations.
- Another example is fitting a line or polynomial to pairs of observations (x_n,y_n) for n = 1,…,N.
- Data is required for estimating the parameters.
- Once estimated, these models are relatively compact (requiring structural form and parameters).
- It is relatively easy to explain and transport these models.

Non-parametric approach

- A non-parametric approach does not make any assumption about the underlying structure that is being approximated.
- This is attractive in situations where we do not know what to expect.
- It may be that neither science nor domain knowledge are sufficient to provide a definitive model structure.
- These models are relatively parsimonious but typically require provision of an entire database.

Types of distributions



Quiz

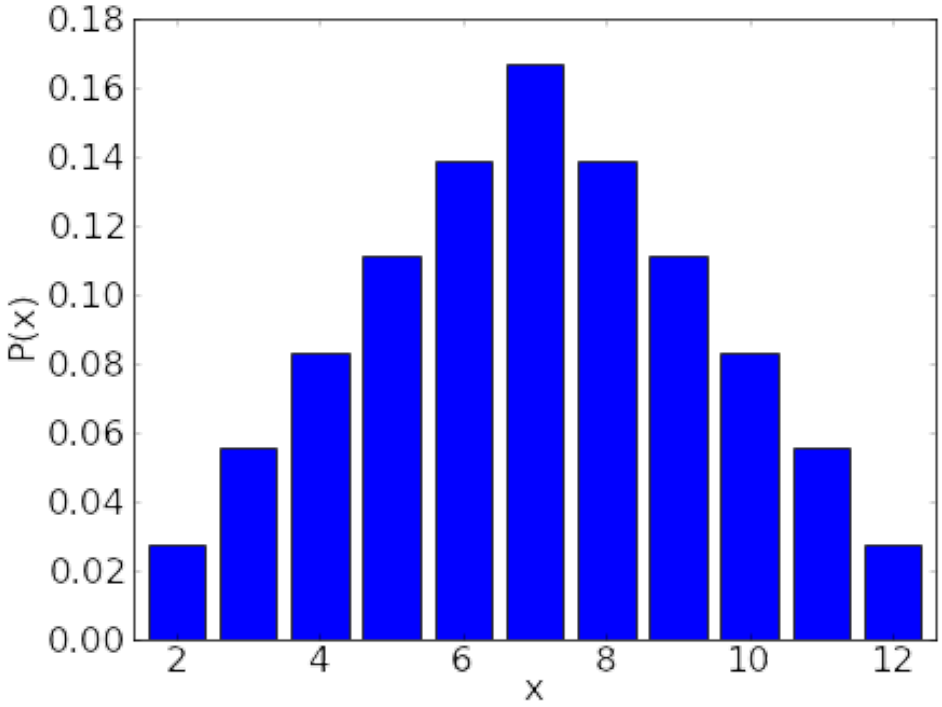
- What is the probability of obtaining a seven from the sum of two six-sided dice?

a) 1/18
b) 1/12
c) 1/9
d) 1/6

- **Slido.com**
- **#81767**

Sum of two six-sided dice

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6



Sum	Probability
2 or 12	1/36
3 or 11	2/36 = 1/18
4 or 10	3/36 = 1/12
5 or 9	4/36 = 1/9
6 or 8	5/36
7	6/36 = 1/6

Traders of Catan Board game. Robber probability is 1/6.

Parametric - Distributions

- Normal distribution - heights
- Log-normal distribution – financial returns
- Poisson distribution – waiting times
- Bernoulli distribution – coin toss
- Student’s t-distribution – hypothesis testing
- Weibull distribution – wind speeds

Your Net Worth Number

Your number	Bottom of your bracket	How many adults in the world are in your bracket	What you can afford	Who
-2	$0.01	1.5 billion	Very little. This category includes people with negative net worth. So you're either poor—or a rich person on a bad day, with liabilities exceeding your assets.	Subsistence farmer
-1	$0.10			
0	$1			
1	$10			
2	$100			
3	$1,000	1.7 billion	Cover small emergency without borrowing	Median American renter
4	$10,000	1.3 billion	New car	Median American family headed by someone with no college education
5	$100,000	436 million	Mortgage	Alexandria Ocasio-Cortez (after a year or two on a congressional salary)
6	$1 million	40 million	Second home by the shore	Boris Johnson
7	$10 million	1.7 million	Third home wherever you want	Ginni Rometty
8	$100 million	49,000	Name on a college building	Rex Tillerson
9	$1 billion	2,700	Name on a college	Silvio Berlusconi
10	$10 billion	150	Sports team in major market	Elon Musk
11	$100 billion	2	Space travel, eradication of polio	Jeff Bezos and Bill Gates. Really, just those two.

Data: Credit Suisse Global Wealth Report 2018 for worth numbers -2 through 8. Bloomberg Billionaires Index for 9-11. Federal Reserve, Financial Samurai, Bloomberg Reporting, Bloomberg Billionaires Index.

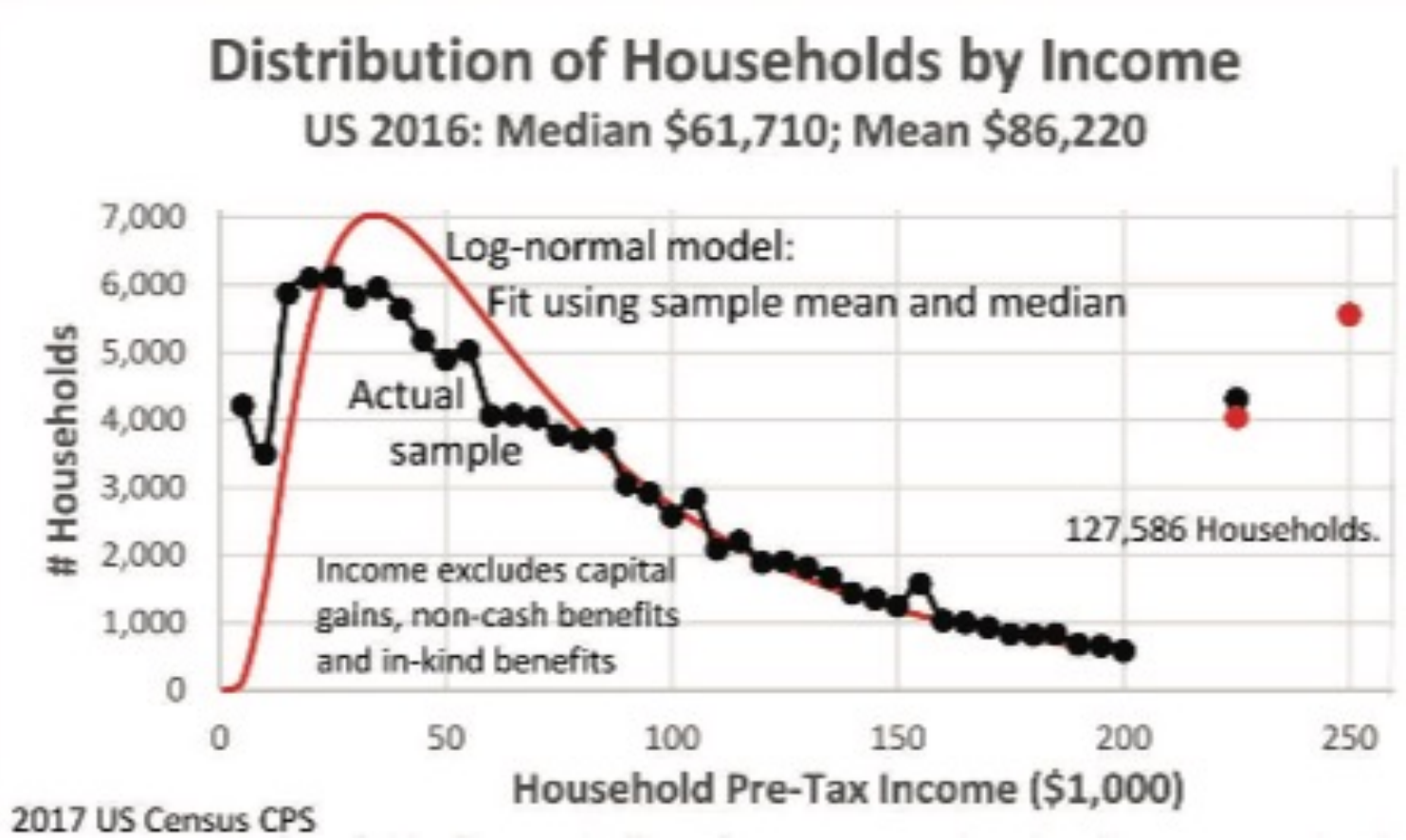
Poll

- What kind of distribution does income typically follow?

a) Normal
b) Log-normal
c) Poisson
d) Binomial

- **Slido.com**
- **#81767**

Household Income



Evolution of median income

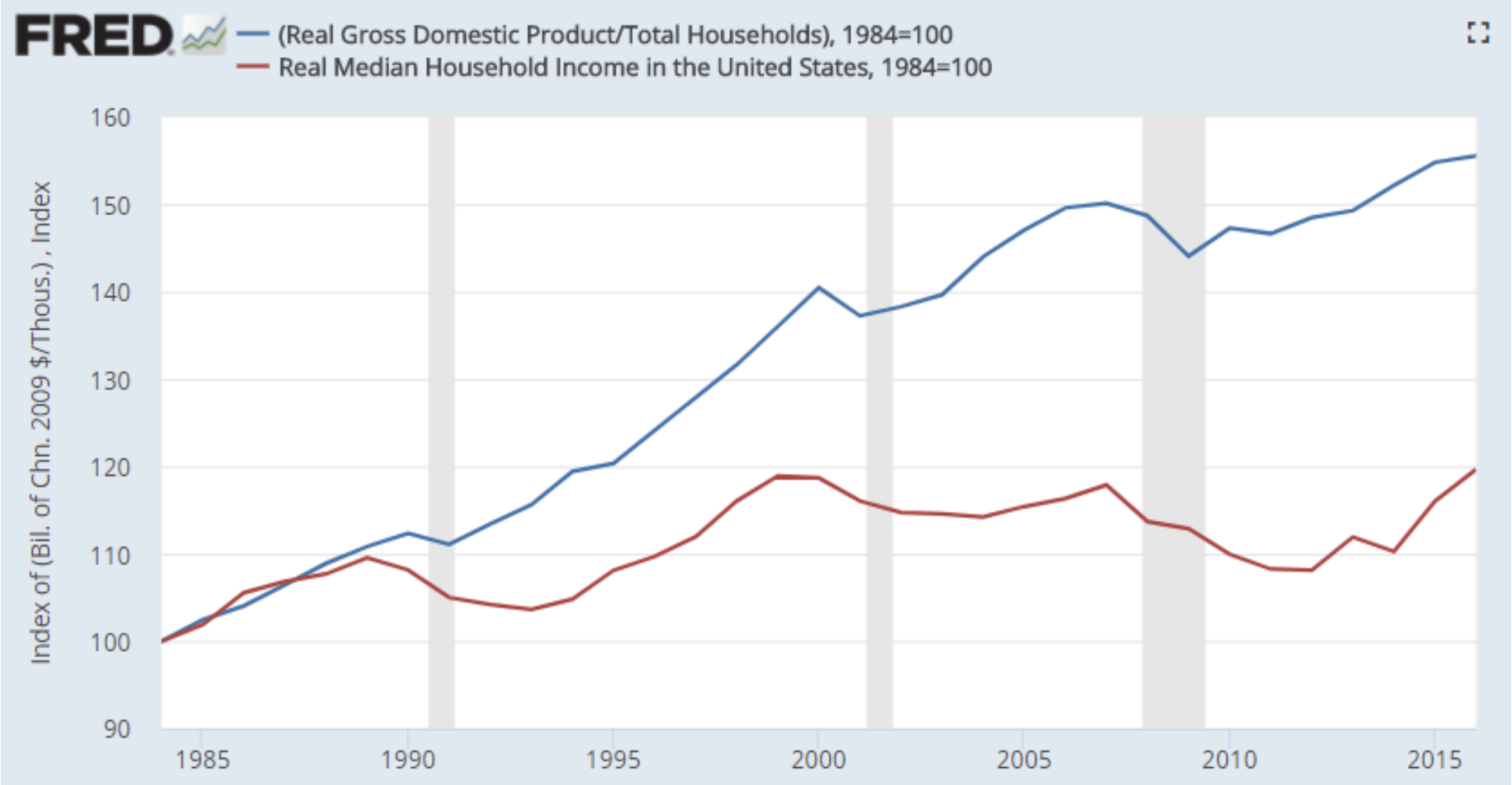
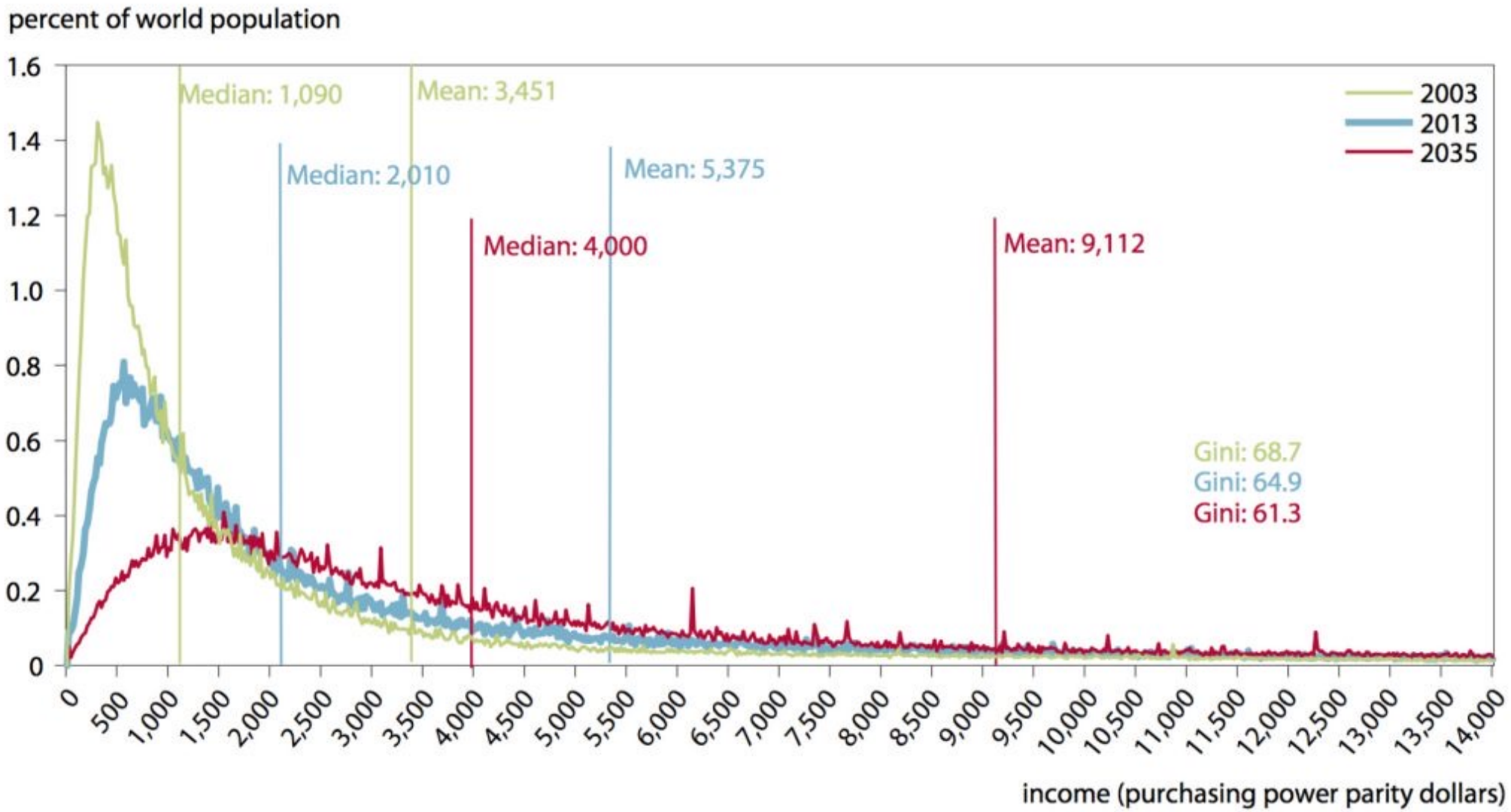


Figure 5 Frequency plot of global income distribution, 2003, 2013, and 2035

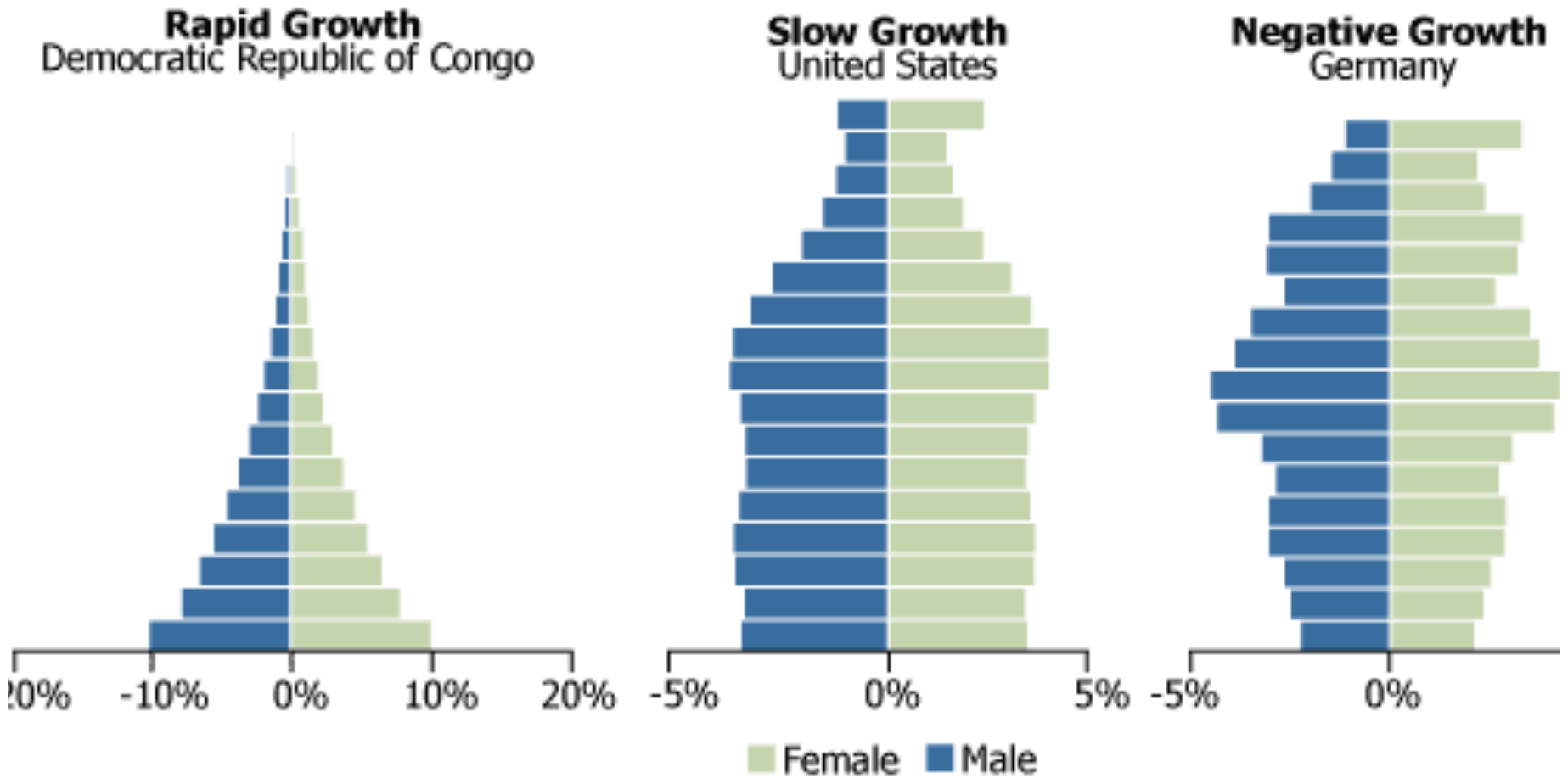


Notes: Percent of world population for each $20 interval is reported on the vertical axis. Individual incomes on the horizontal axis are expressed in US dollars at 2011 international prices (purchasing power parity)

Sources: OECD, Consensus Forecasts, IMF/World Bank, and authors' forecasts for growth; United Nations for population projections; Luxembourg Income Study and World Bank for household survey data on income distribution.

https://www.gapminder.org/data/documentation/income-mountains-dataset/

Age Histograms



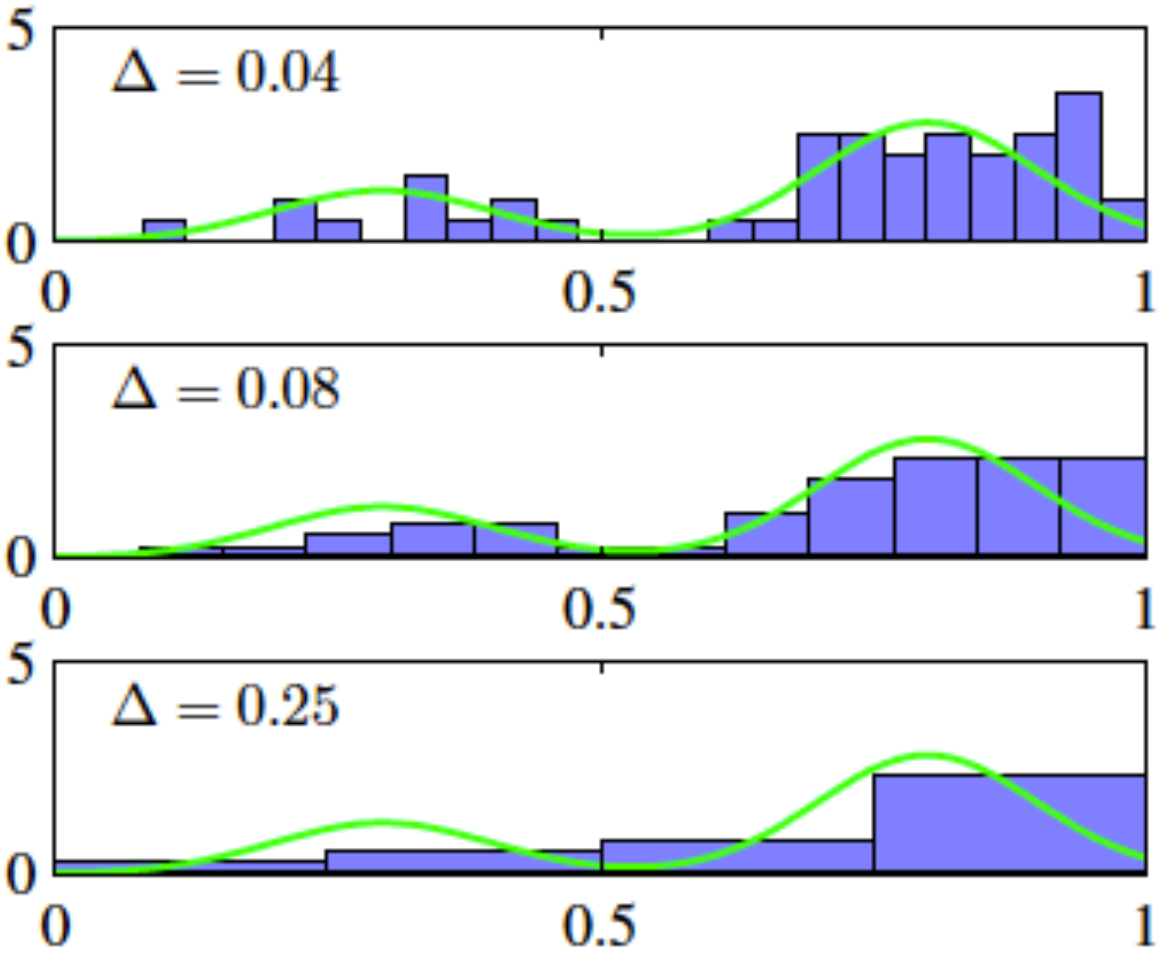
Source: UN World Population Prospects, The 2006 Revision.

Density estimation from histogram

- Consider a histogram for a one-dimensional variable x.
- A histogram is typically constructed by simply partitioning *x* into distinct bins of width Δ.
- If n_i observations fall into the ith bin, then the density estimate for x is calculated by normalizing by the total number of observations and the size of the bins:

$$p_i = n_i/(N\Delta)$$

Bin size effect



Source: Chris Bishop (2006)

Kernel density estimation

- An estimate of the density at x in a D dimensional space is given by

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h^D}K_h(x - x_n)$$

- where h is a bandwidth and $K_h(u)$ is a kernel function satisfying:

$$K_h(u) \geq 0$$

$$\int K_h(u)\,du = 1.$$

Kernels

- The general idea of a kernel is to give most weight to nearby points and less weight to points that are further away.
- A simple window or hat kernel is given by

  $K_h(u) = 1$ if $|u| \leq h$

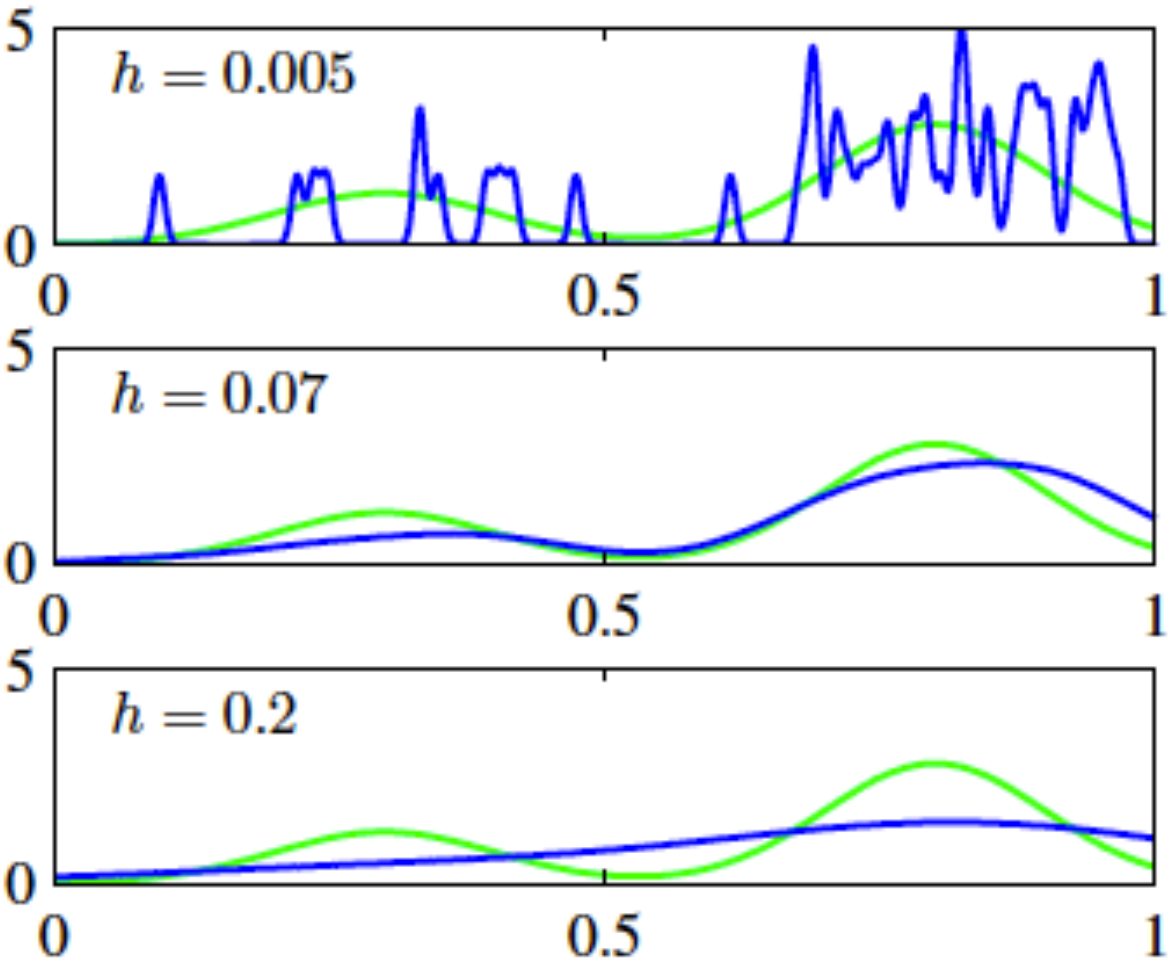
  $K_h(u) = 0$ otherwise
- This particular kernel has an abrupt change in weight at the boundaries.

Gaussian Kernel

- To avoid discontinuities, it may be better to select a smoother kernel function.
- This should provide a smoother density model.
- The Gaussian kernel is a popular choice:

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\sqrt{2\pi h^2}}\exp\left(\frac{\|x - x_n\|^2}{2h^2}\right)$$

Example of kernel density estimation



Source: Chris Bishop (2006)

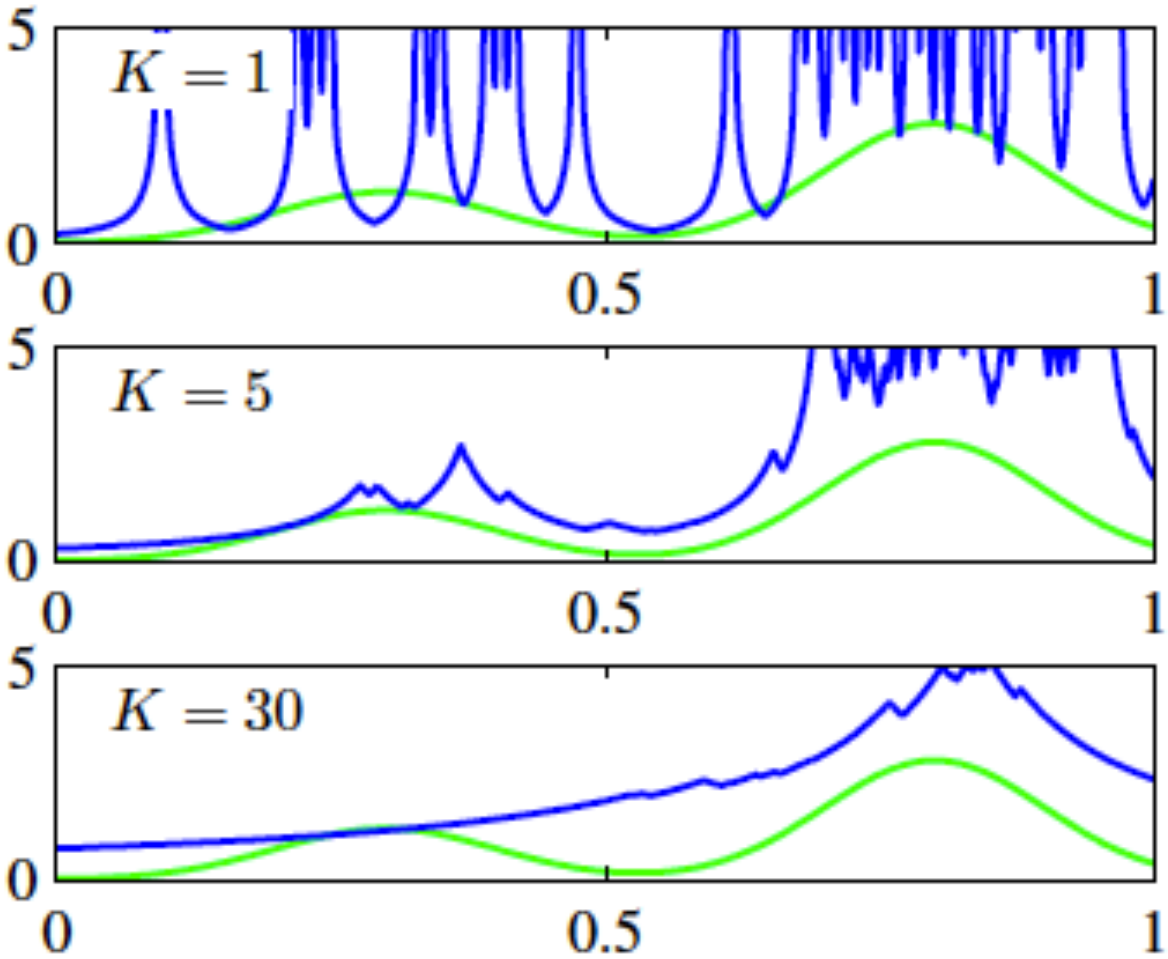
K nearest neighbors (KNN)

- A KNN density estimate is given by

  $$p(x) = K/NV$$

- where N is the total number of data points, and V is the volume required to include the K nearest neighbors.
- The idea is to fix K in advance and then determine V by considering a neighborhood around x, defined by a sphere, that grows until it captures the K nearest neighbors.

KNN density estimation



Source: Chris Bishop (2006)

Poll

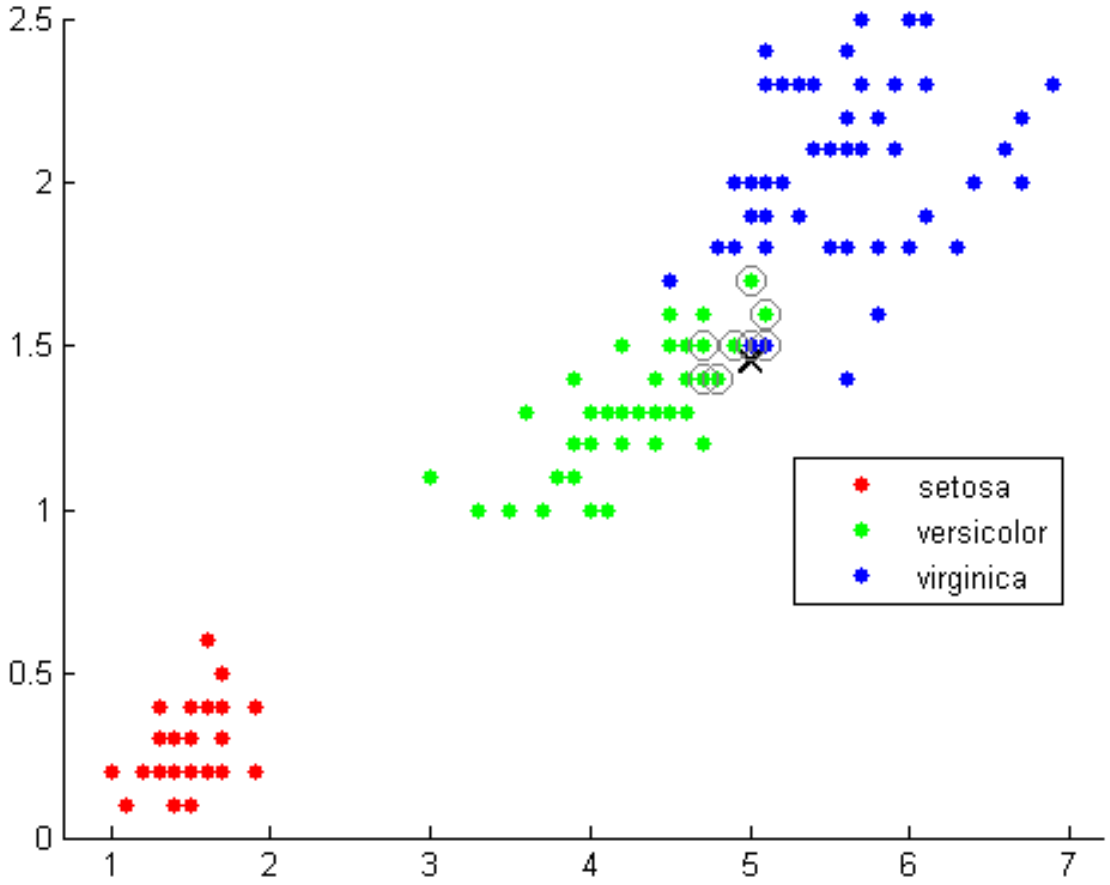
- k nearest neighbours (KNN) techniques are sensitive to the units for the variables but not the distance metric selected.
  - True
  - False

- **Slido.com**
- **#81767**

KNN classification

- KNN can also be used to provide a non-parametric approach for classification.
- Given a test point x, find the K nearest neighbors from the training data set.
- If K_k points belong to class C_k, then the probability that x belongs to class C_k is given by $p(C_k|x) = K_k/K$.
- The test point x is then classified as belonging to the class with the largest value of $p(C_k|x)$.

KNN



K nearest neighbour (KNN) classifies a new point based on the labels given to its k nearest neighbours.

KNN

- Unlike some other approaches, the KNN model gives equal importance to all input variables.
- Therefore relevant input features should be selected before using the KNN.
- There are three components that define a KNN model:
  - the number of nearest neighbors;
  - the distance metric; and
  - the weighting system.

KNN – Weights

- A fixed number of neighbors will define neighborhoods of different sizes depending on the local density of the data.
- Weights provide a means of allowing nearby points to have more influence on the estimate.
- The weights usually decay with distance from the input.
- A weighting system can be applied to all points or only those in a local neighborhood of the input.

KNN – distance metrics

- Distance metrics include the following:
- Euclidean distance
- Standardized Euclidean distance - each coordinate difference between rows in X and the query matrix is scaled by dividing by the corresponding element of the standard deviation.
- Mahalanobis distance – standardized distance computed using a positive definite covariance matrix.
- Chebychev distance - maximum coordinate difference.
- Correlation - one minus the sample linear correlation between observations.

Kernel regression

- Kernel regression is a non-parametric technique in statistics to estimate the conditional expectation of a random variable.
- The objective is to find a non-linear relation between a pair of random variables X and Y.
- The idea is that a kernel can be used to provide a smooth representation of this relationship.

Kernel Methods

- Given pairs of observations (x_n, y_n) we can also apply kernel density estimation to obtain an approximation to the relationship between x and y.
- We wish to determine an estimate of y for any given x.
- Like KNN, the kernel can be used to ensure that only nearby values of x_n and their associated y_n contribute to the estimate of y.
- This is achieved by centering a kernel on each observation x_n and using these to provide weights for estimating y:

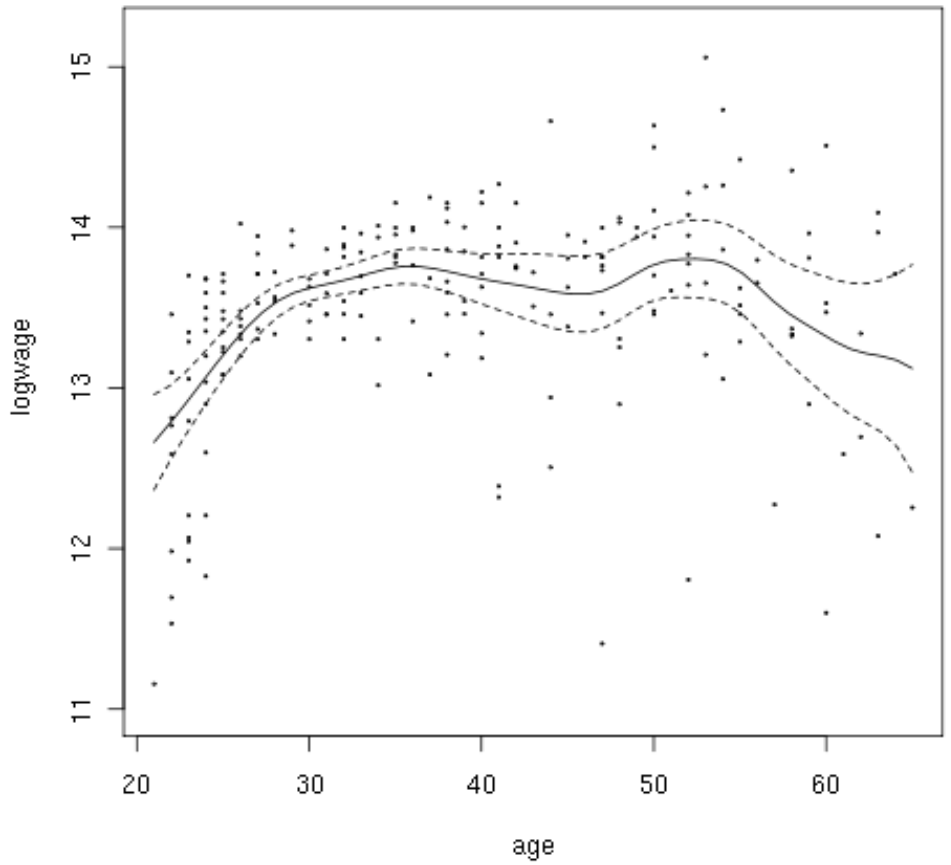
$$E(y \mid x) = \frac{1}{N} \sum_{n=1}^{N} w_n y_n$$

Nadaraya-Watson estimator

- Consider pairs of observations (x_n, y_n) such that $y_n = m(x_n) + \varepsilon(x_n)$ where g is the underlying signal and ε is noise.
- We can obtain an estimate of m(x) given by

$$m(x) = E(y \mid x) = \frac{\sum_{n=1}^{N} K_h(x - x_n) y_n}{\sum_{n=1}^{N} K_h(x - x_n)}$$

Example: Wage versus Age

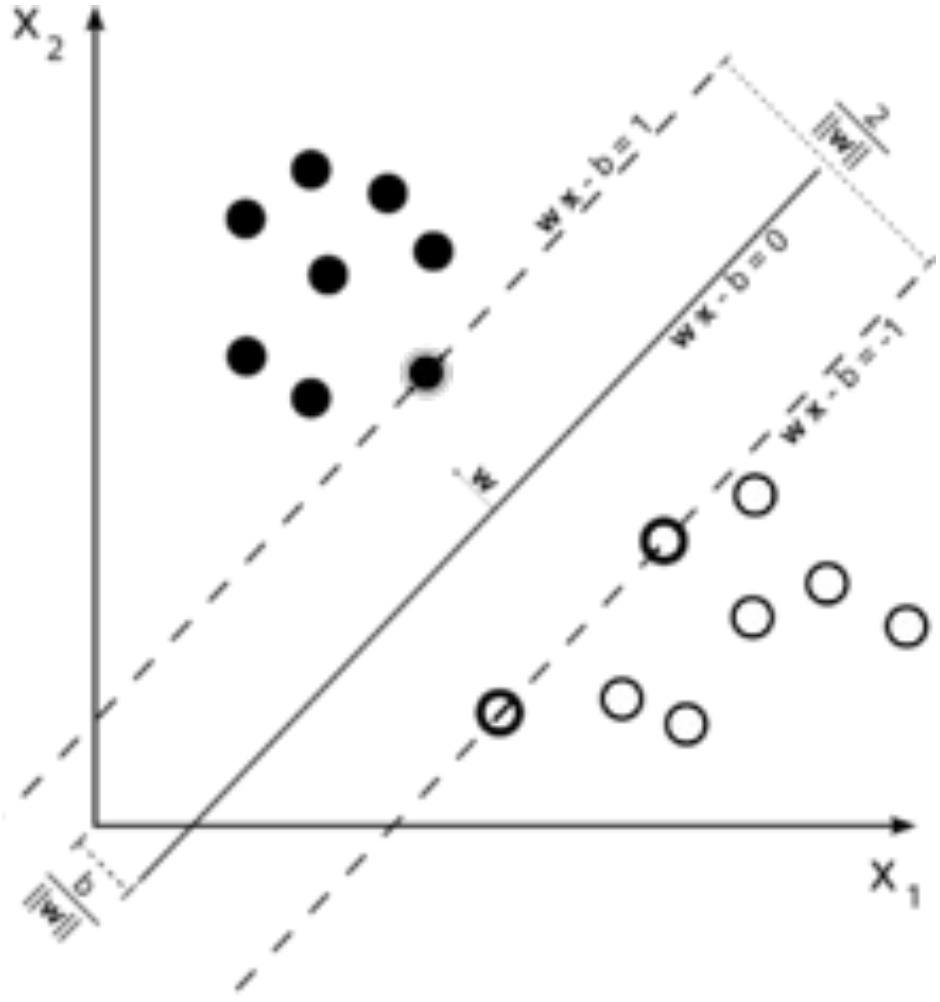


Source: Canadian wages (Wikipedia).

Support Vector Machines (SVM)

- Support Vector Machines (SVMs) provide a means of binary classification.
- The SVM model assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.
- The SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.
- New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

SVM

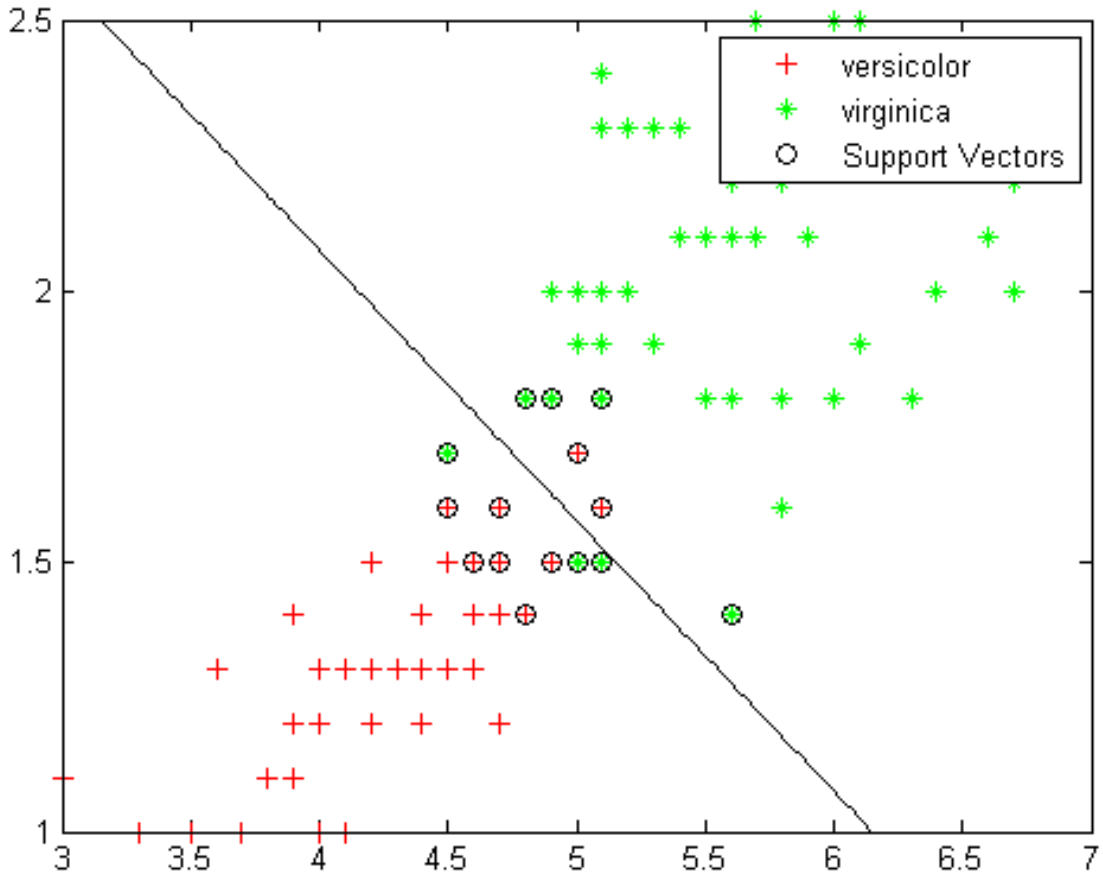


Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

SVM optimization

- Points are labeled as $y_i = 1$ or $y_i = -1$.
- Any hyperplane: $\mathbf{w}.\mathbf{x} - b = 0$.
- Margin hyperplanes: $\mathbf{w}.\mathbf{x} - b = \pm 1$.
- Distance between hyperplanes is $2/||\mathbf{w}||$.
- We want $y_i(\mathbf{w}.\mathbf{x}_i - b) \geq 1$ to ensure points are classified correctly.
- Minimize $||w||$ subject to $y_i(\mathbf{w}.\mathbf{x}_i - b) \geq 1$.

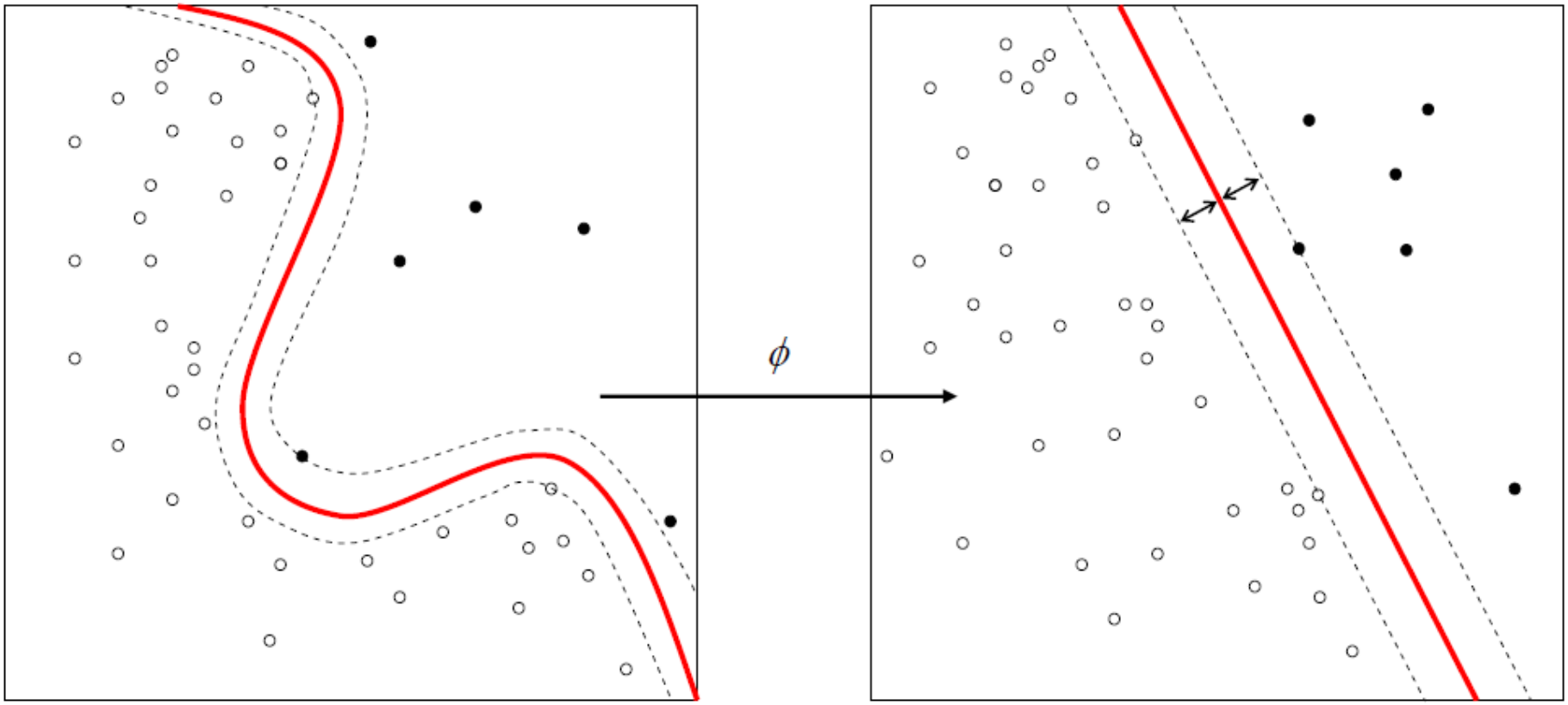
Support Vector Machines (SVM)



Nonlinear SVM

- Nonlinear SVM classifiers are constructed by applying a kernel trick to maximum-margin hyperplanes.
- The resulting algorithm is formally similar, except that every dot product is replaced by a nonlinear kernel function.
- This allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space.
- The transformation may be nonlinear and the transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional feature space, it may be nonlinear in the original input space.

Nonlinear SVM



Source: wikipedia

Matlab: KNN regression

- Learning: {Xl,yl}
- Testing: {Xt,yt}

- K = 2.^[0:5];
- for k=1:length(K)
- [idx, dist] = knnsearch(Xl,Xt,'dist','seuclidean','k',K(k));
- ythat = nanmean(yl(idx),2);
- E = yt - ythat;
- RMSE(k) = sqrt(nanmean(E.^2));
- end

Q&A