Introduction to data mining

MATH 6312
Department of mathematics, UTA

Course information

- Course Webpage: find a link in Blackboard.
- Office Hours: SEIR 218, Tu/Thu 4:30-5:30 pm or by appointment
- Prerequisite: Basic programming skills are preferred, but not required.
- Required Textbooks
  - Hastie, T., Tibshirani, R., and Friedman, J. H. (2008), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer.
- Other Recommended Textbooks and Resources
  - Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
  - Rencher and Schaalje (2008). Linear Models in Statistics, 2nd Edition. Wiley.

Assignments

- Midterm take-home exam (20%)
  - The exam will be distributed during the class. Students need to submit the solution within 48 hours.
- Homework assignments (40%)
  - 7 HW assignments (30%) and final HW (10%). The lowest homework score will be dropped, but you cannot drop the final homework.
- Final projects (40%)
  - The final project guideline will be announced later.

Assignment submission format

- All assignments (homework, midterm exam, final project) must be turned in electronically, through Blackboard.

- All assignments (unless stated otherwise) must be submitted in R Markdown and PDF format.

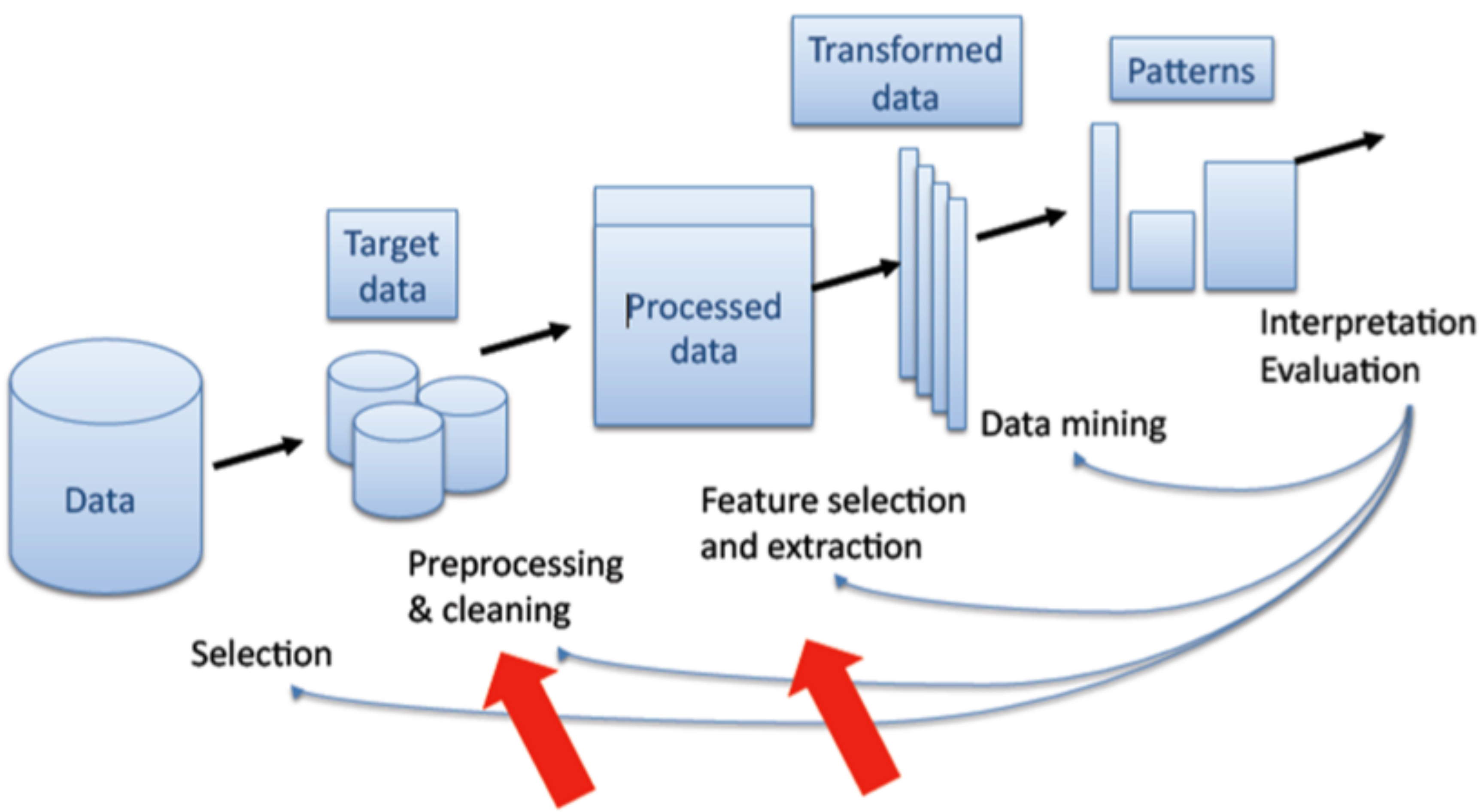
- Work submitted in R Markdown format that does not compile, i.e., fails "Knit PDF (or HTML)", will receive an automatic grade of 0.

What is data mining?

- Many definitions
- Non-trivial extraction of implicit, previously unknown and useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

What is data mining?

- Data mining is actually a part of the knowledge discovery process (KDD: knowledge discovery from data).



Why data mining?

- Data collected and stored at enormous speeds (GB/hour)
  - Remote sensors on a satellite
  - Genomic data
  - Web data, e-commerce
  - Bank/credit card transactions
- Computers have become cheaper and more powerful.
- Traditional techniques infeasible for raw data.

Origins of data mining

- Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
    - Scalability of number of features and instances
    - Stress on algorithms whereas foundations of methods and formulations provided by statistics and machine learning.

Data mining tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.
- Classification, clustering, regression, anomaly detection, etc.

Email spam data

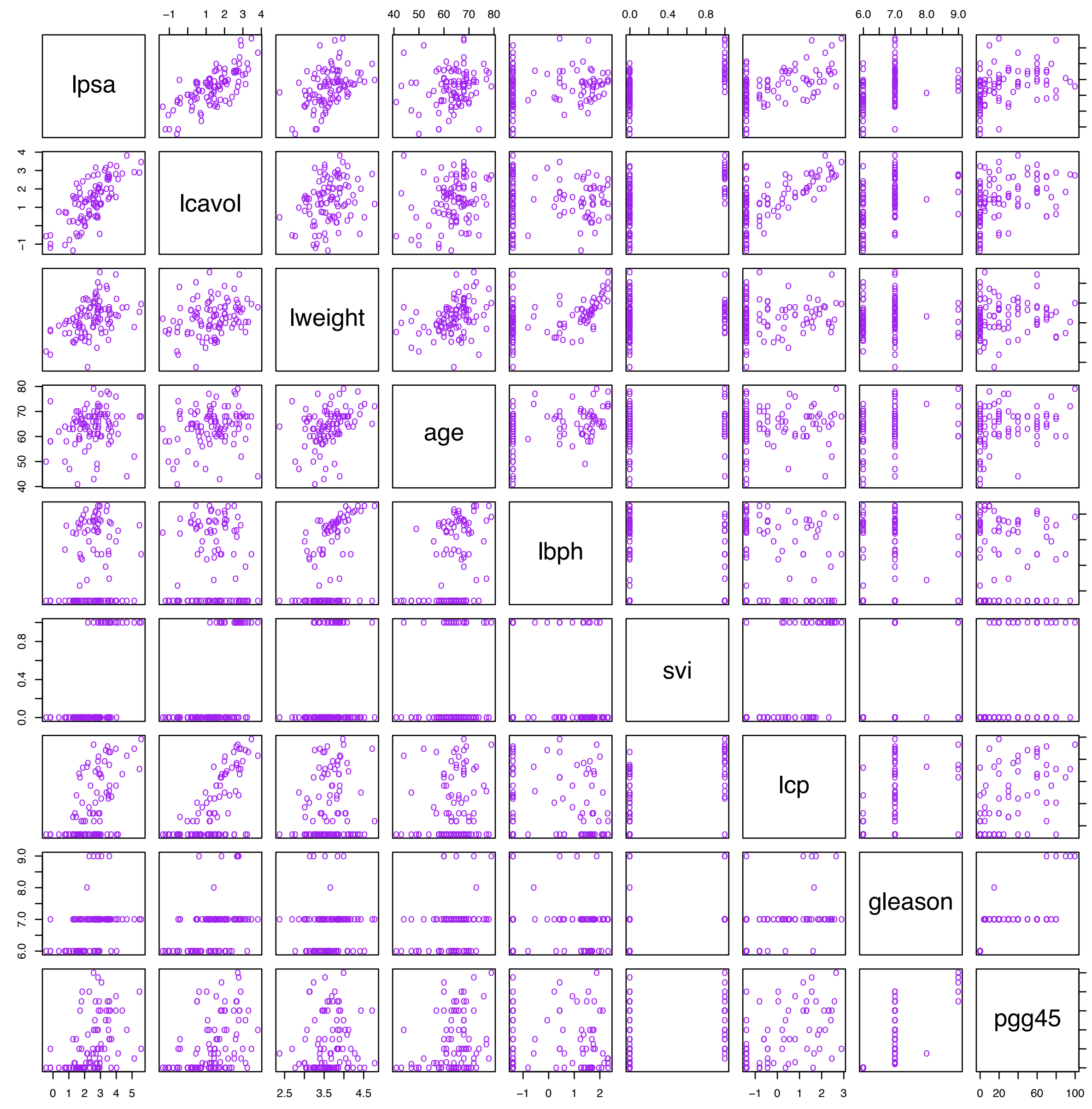
- A goal is to predict whether the email was spam or not.
- The true outcome (ham or spam) is available, along with the relative frequencies of 57 of the most commonly occurring words and punctuation marks in the email message.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

- A learning method has to decide which features to use and how: for example,
- $$\text{if } (\texttt{\%george} < 0.6)\ \&\ (\texttt{\%you} > 1.5) \quad \text{then } \texttt{spam} \quad \text{else } \texttt{email}.$$

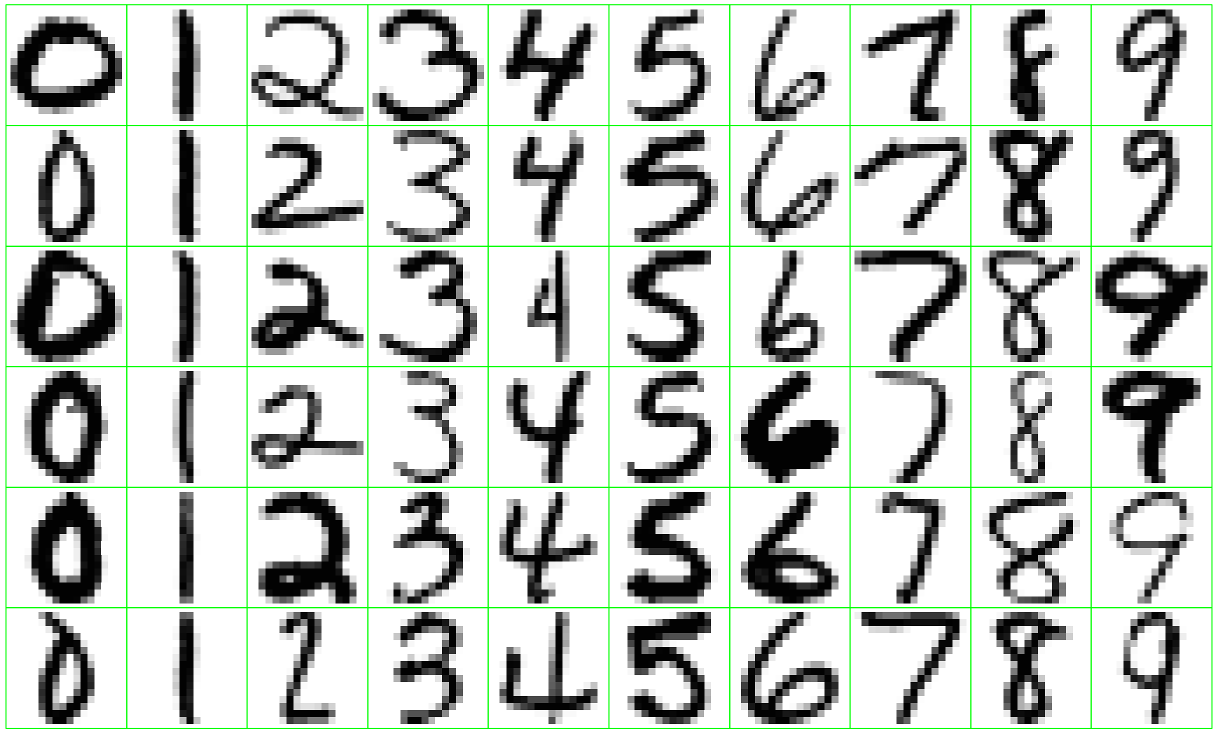
Prostate cancer

- The elevated log of prostate specific antigen (lpsa) may indicate prostate cancer.
- The goal is to predict ipsa from a number of measurements.
- From the scatterplot matrix of the variables, some correlations with lpsa are evident, but a good predictive model is difficult to construct by eye.



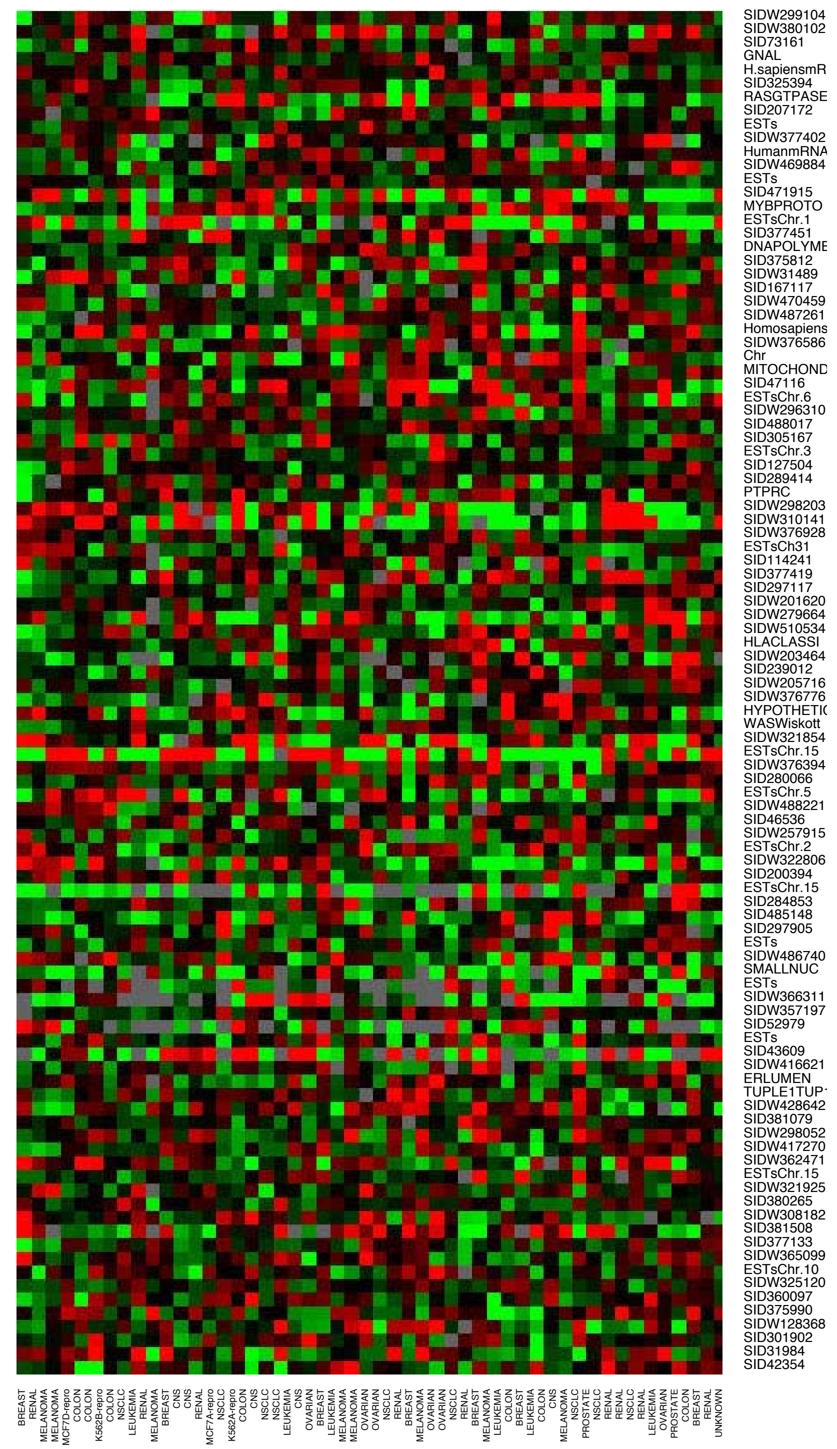
Handwritten digit recognition

- Handwritten ZIP codes on envelopes from U.S. postal mail.
- 16×16 eight-bit grayscale maps, with each pixel ranging in intensity from 0 to 255.
- It is very important to predict, from the 16 × 16 matrix of pixel intensities, the identity of each image (0, 1, . . . , 9).
- The error rate needs to be kept very low to avoid misdirection of mail



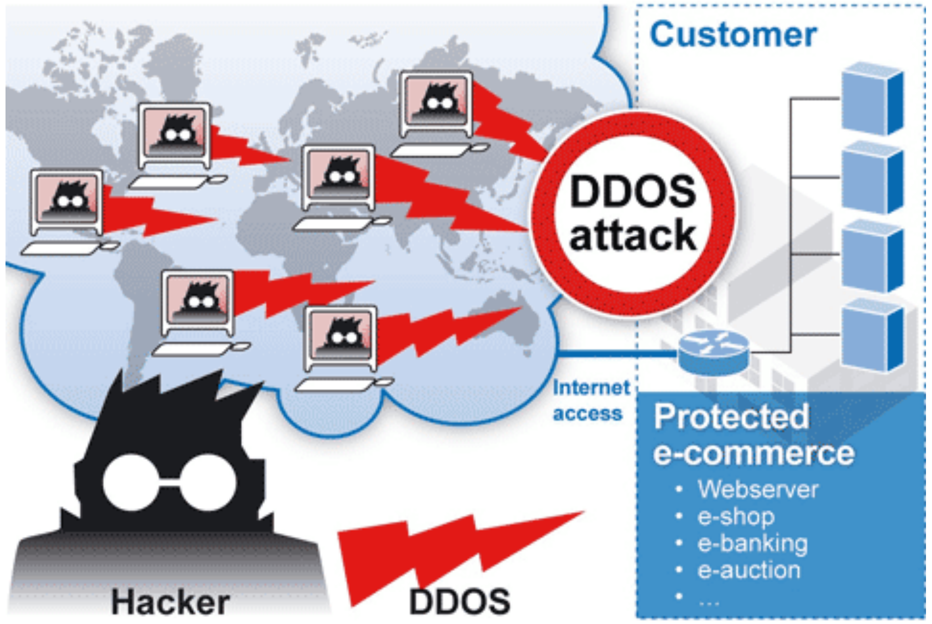
Gene expression data from DNA microarray

- Columns - genes; Rows - samples (tumor types)
- Ranging from bright green (negative, under expressed) to bright red (positive over expressed).
- How the genes and samples are associated?
  - Which samples are most similar to each other, in terms of their expression profiles across genes?
  - Which genes are most similar to each other, in terms of their expression profiles across samples?
  - Do certain genes show very high (or low) expression for certain cancer samples?



Anomaly detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



Characteristics of data in data mining

- Large n (number of observations) and large p (number of predictors); Predictor variables are of mixed types — "curse of dimensionality" (Bellman, 1957)
- Mostly coming from observational studies, instead of designed experiments. Data are constantly accumulating and dynamically varying.
- Data cleaning and preparation (missing value handling, outliers, categorical variables, etc.)
- Exploratory data analysis is vital; may use 90% of the total analysis time.

Supervised vs unsupervised

- Supervised Learning (Learning with a teacher): predictive modeling, i.e., predicting one (or more) output (or target) variable from a set of inputs or predictors; classification (also called pattern recognition) and regression.

- Unsupervised learning (learning without a teacher): Clustering; Segmentation; Dimension reduction; Exploring associations; etc.

- Semi-supervised learning: The training sample contains both labelled and unlabelled data, typically a small amount of labelled data with a large amount of unlabelled data.

Supervised learning

- One response (also called dependent variable, target, outcome, output) Y vs a set of p predictors (also called inputs, independent variables, covariates, features).
  - In the regression problem, Y is quantitative (e.g. price, blood pressure)
  - In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
  - Predictors could be of mixed types: nominal, ordinal, interval, and ratio.

Supervised learning

- Data typically consist of n iid observations:

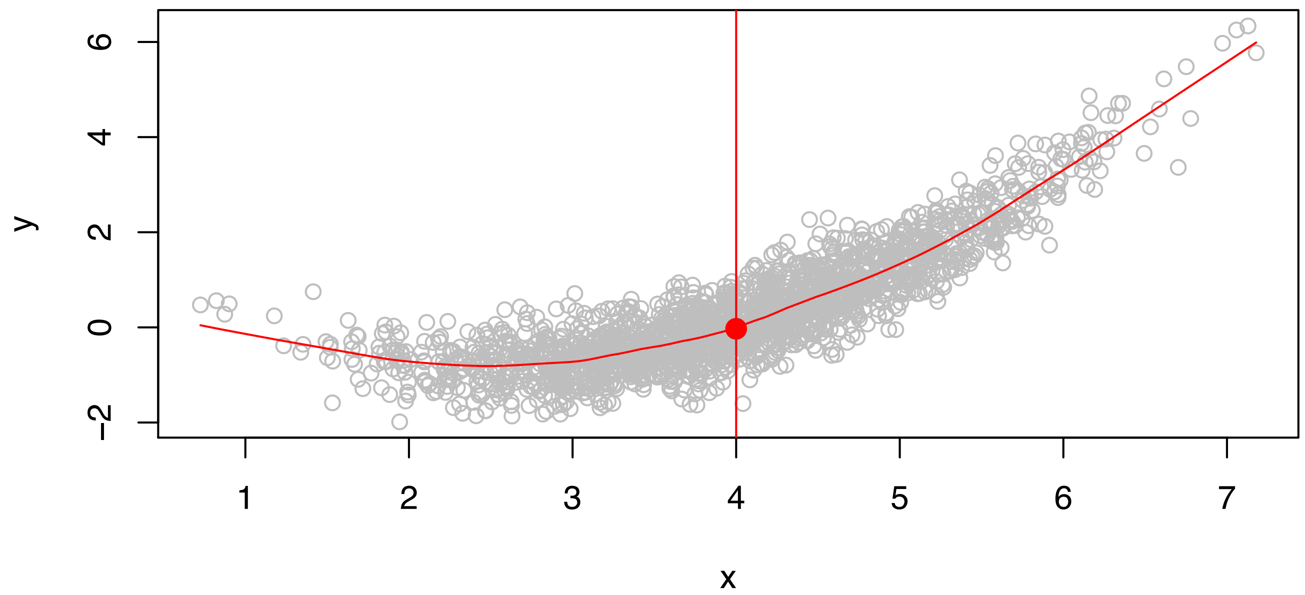
$$\{(y_i, x_{i0}, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$$

- Predictive Modeling — seek a model to predict Y

- The model is a learner, which provides an answer. This answer will be evaluated by comparing with the corrected answer or the target (i.e., the observed value provided by the teacher)

Regression

- What is a good model f(x) to make predictions of Y at X=x?



- There can be many Y values at X=4. A good value is

$$f(4) = E(Y|X = 4)$$

- This ideal f(x) is called the regression function.

Regression

- The regression function is optimal predictor of Y with regard to expected prediction error under squared loss function.

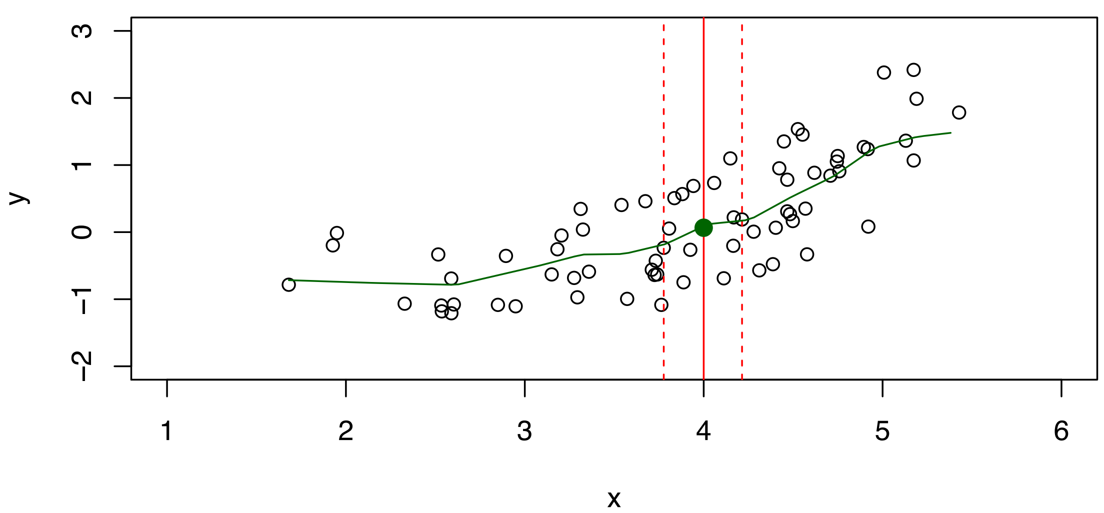
  - Loss function (squared loss):
  $$L(Y, f(x)) = (Y - f(X))^2$$
  - Risk function:
  $$E(L(Y, f(X)) = E(Y - f(X))^2$$
  - By choosing $f(X) = E(Y|X)$, the expected prediction error is minimized.
  $$E(Y - f(X))^2 = E_X E([Y - f(X)]^2|X)$$

How to estimate f(X)?

- Typically we have few data points with X = 4. So it is hard to obtain the regression function.

  - Nearest neighbor method: relax the definition and let

$$\hat{f}(x) = \text{Ave}(Y|X \in N(x))$$

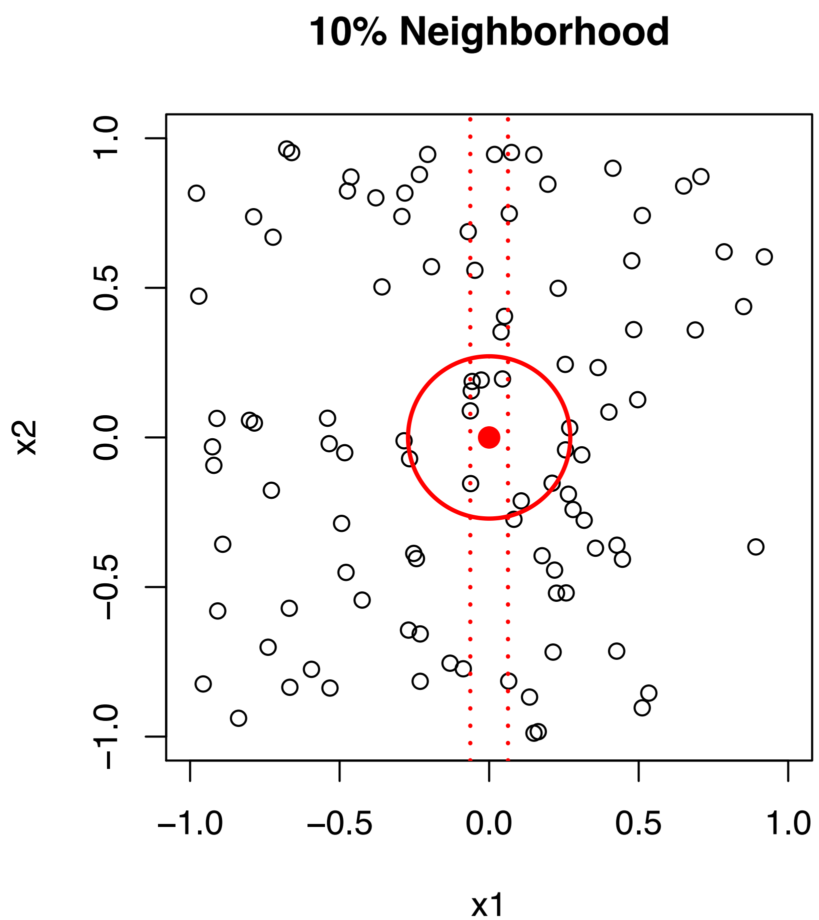


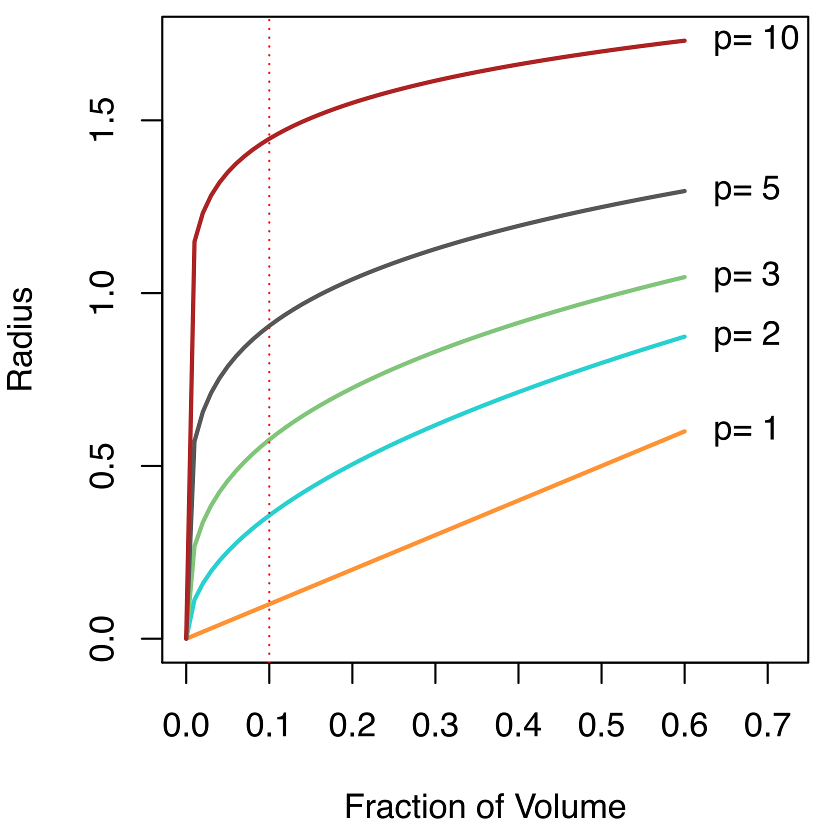
  - Linear regression: assume the regression function is linear.

$$E(Y|X) = \alpha + \beta X$$

Curse of dimensionality

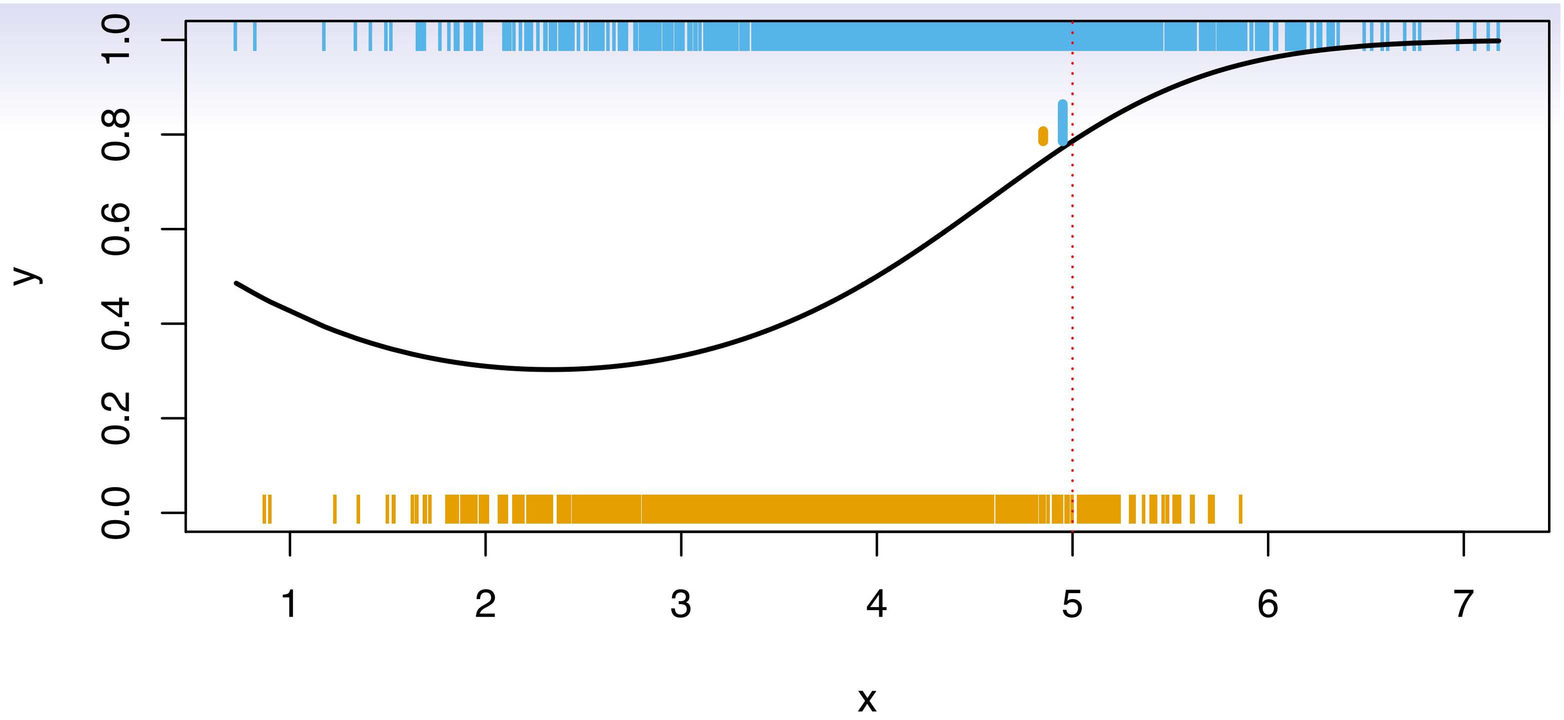
- Nearest neighbor methods can be lousy when p is large.
  - K observations, say 10% of observations, that are nearest to a given X=x may be very far away from x when p is large, leading to a very poor prediction.





Classification

- The response variable Y is qualitative — e.g. email is one of (spam,ham) (ham = good email), digit class is one of $\{0,1,\ldots,9\}$. Our goals are to
  - Build a classifier $C(X)$ that assign a class label to a future unlabeled observation.
  - Assess the uncertainty in each classification.
  - Understand the roles of the predictors



- Suppose there are two classes (Y = 0 or 1).
- Let $p_k = \Pr(Y = k|X), K = 0, 1$
- The Bayes optimal classifier is
- $$C(X) = j \text{ if } p_j(X) = \max\{p_0(X), p_1(X)\}$$

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well your are doing.
- Different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Unsupervised learning

- We observe only the features.

$$\{(x_{i0}, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$$

- We are not interested in prediction, because we do not have an associated response variable Y.
  - Principal components analysis: a tool used for data visualization or data pre-processing before supervised techniques are applied.
  - Clustering: a broad class of methods for discovering unknown subgroups in data. .

Advantage of unsupervised learning

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.

- But techniques for unsupervised learning are of growing importance in a number of fields:

  - Subgroups of breast cancer patients grouped by their gene expression measurements.

  - Movies grouped by the ratings assigned by movie viewers.

- It is often easier to obtain unlabeled data — from a lab instrument or a computer — than labeled data, which can require human intervention.

Statistical learning vs machine learning

- Machine learning arose as a subfield of artificial intelligence.
- Statistical learning arose as a subfield of statistics.
- There is much overlap — both fields focus on supervised and unsupervised problems.
- Machine learning has a greater emphasis on large scale applications and prediction accuracy.
- Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.

Challenges of data mining

- Scalability
- Dimensionality
- Complex and heterogeneous data
- Data quality
- Data ownership and distribution
- Streaming data

Meaningfulness of answers

- A big data-mining risk is that you will "discover" patterns that are meaningless.
- Rhine Paradox: a great example of how not to conduct scientific research.
  - Joseph Rhine was a parapsychologist in the 1950's who hypothesized that some people had extra-sensory perception (ESP).
  - He devised (something like) an experiment where subjects were asked to guess 10 hidden cards (red or blue). He discovered that almost 1 in 1000 had ESP. They were able to get all 10 right!
  - He told these people they had ESP and called them in for another test of the same type. Alas, he discovered that almost all of them had lost their ESP.
  - He concluded that you shouldn't tell people they have ESP; it causes them to lose it.