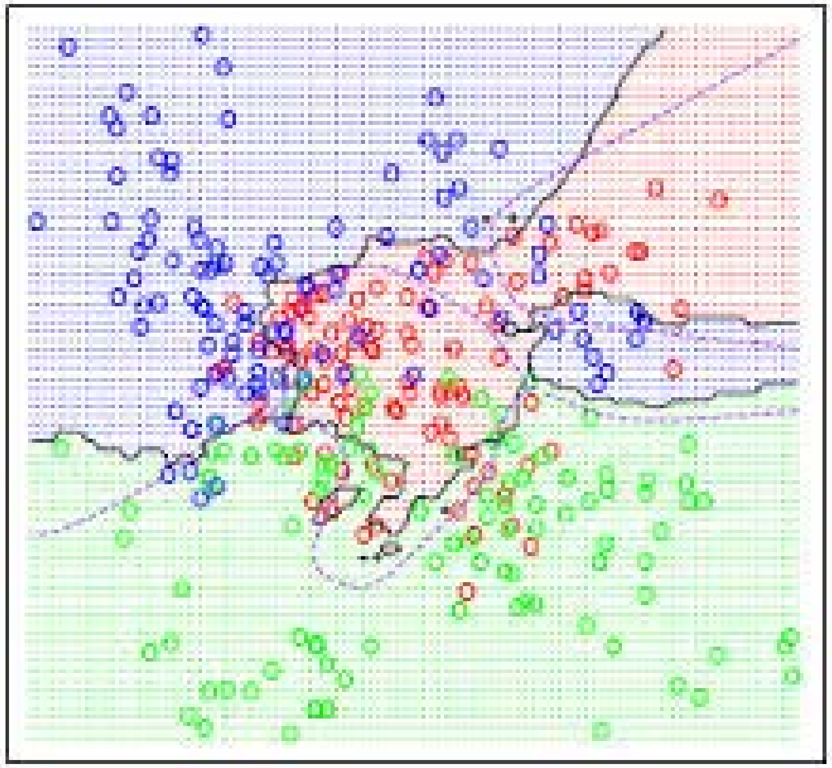


COMP90042 LECTURE 2

TEXT CLASSIFICATION

OUTLINE

- Fundamentals of classification
- Algorithms for classification
- Text classification tasks

CLASSIFICATION

- Input
  - A document d
    - Often represented as a vector of *features*
  - A fixed output set of classes $C = \{c_1, c_2, \ldots c_k\}$
    - Categorical, not continuous (regression) or ordinal (ranking)
- Output
  - A predicted class $c \in C$

TYPES OF CLASSIFICATION

- Rule-based vs. statistical
  - Rule-based methods often accurate on simple problems
  - Statistical models usually preferred for large feature spaces
- Supervised vs. unsupervised
  - Supervised usually more accurate
  - But generally requires hand-labelled data
- Generative vs. discriminative
  - Generative classifiers model how classes generate features
  - Discriminative classifiers find best discriminating features

BUILDING A TEXT CLASSIFIER

1. Identify a task of interest
2. Collect an appropriate corpus
3. Carry out annotation
4. Select features
5. Choose a machine learning algorithm
6. Tune hyperparameters using held-out development data
7. Repeat earlier steps as needed
8. Train final model
9. Evaluate model on held-out test data

EVALUATION: ACCURACY

	Classified As	
Class	A	B
A	79	8
B	13	11

Accuracy = correct classifications/total classifications

= (79 + 11)/(79 + 11 + 13 + 8)

= 0.81

0.81 looks good, but most common class baseline accuracy is 0.83

EVALUATION: PRECISION & RECALL

	Classified As	
Class	A	B
A	79	8
B	13	11

False Positives (fp)

True Positives (tp)

False Negatives (fn)

B as “positive class”

Precision = correct classifications of B (tp) /total classifications as B (tp + fp)

= 11/(11 + 8)

= 0.57

Recall = correct classifications of B (tp)/total instances of B (tp + fn)

= 11/(11 + 13)

= 0.46

EVALUATION: F(1)-SCORE

- Harmonic mean of precision and recall

$$F1 = 2\ \text{precision}*\text{recall}/(\text{precision} + \text{recall})$$

- Like precision and recall, defined relative to a specific positive class
- But can be used as a general multiclass metric
  - Macroaverage: Average F-score across classes
  - Microaverage: Calculate F-score using sum of counts

HYPERPARAMETER TUNING

- Dataset for tuning
  - Development set
  - Not the training set or the test set
  - k-fold cross-validation
- Specific hyperparameters are classifier specific
  - E.g. tree depth for decision trees
- But many hyperparameters relate to regularization
  - Regularization hyperparameters penalize model complexity
  - Used to prevent overfitting
- For multiple hyperparameters, use grid search

CHOOSING A CLASSIFICATION ALGORITHM

- Bias vs. Variance
- Feature independence
- Feature scaling
- Complexity
- Speed

NAÏVE BAYES

- Finds the class with the highest likelihood under Bayes law
  - i.e. probability of the class times probability of features given the class

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

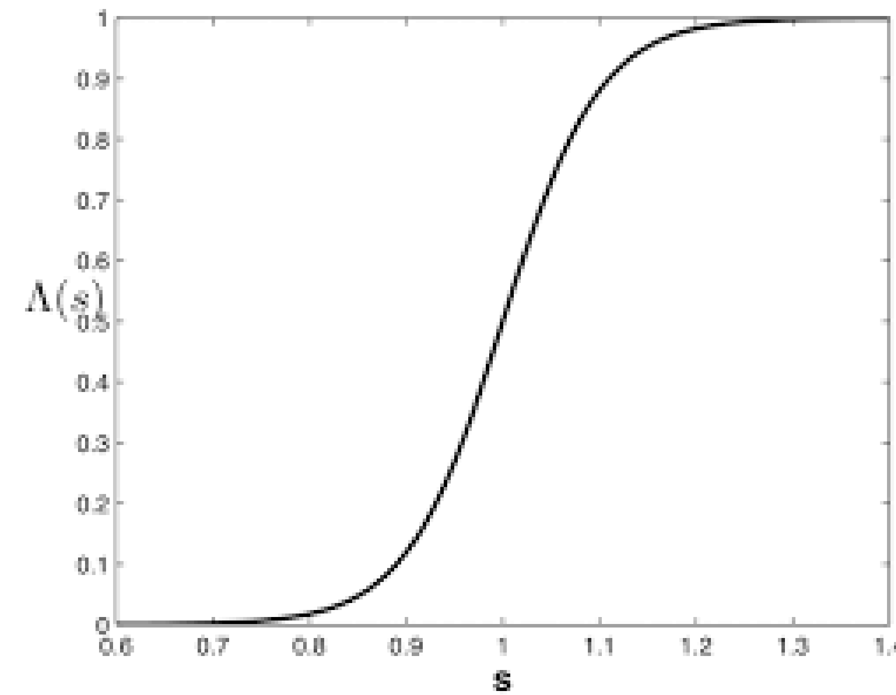
- Naïvely assumes features are independent

$$p(c_n|f_1 \dots f_m) = \prod_{i=1}^{m} \underbrace{p(f_i\,|c_n)}_{\text{The proportion of features (i.e. words) for documents of class } n \text{ which are feature (word) } i} \overbrace{p(c_n)}^{\text{The proportion of training documents which are class } n}$$

NAÏVE BAYES

- Pros: Fast to "train" and classify; robust, low-variance; good for low data situations; optimal classifier if independence assumption is correct
- Cons: Independence assumption rarely holds; low accuracy compared to similar methods in most situations; smoothing required for unseen class/feature combinations

LOGISTIC REGRESSION



- A classifier, despite its name
- Discriminative model: models P(c | d) directly, no need for P(d | c)
- A linear model, but uses *softmax* "squashing" to get valid probability

$$p(c_n | f_1 \ldots f_m) = \frac{1}{Z} \cdot \exp\left(\sum_{i=0}^{m} w_i f_i\right)$$

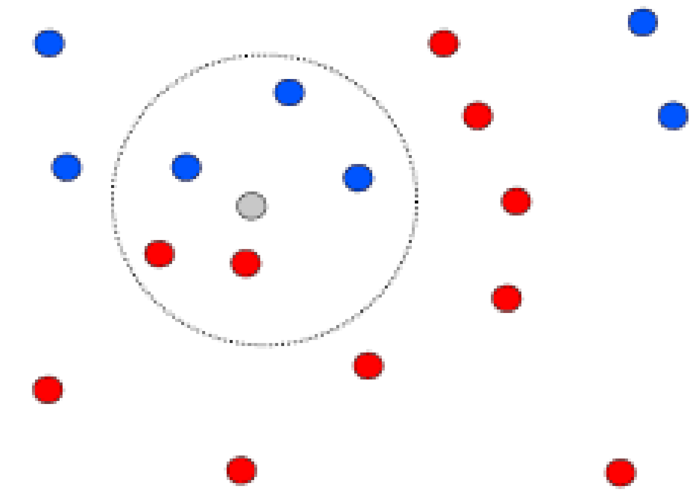
- Training maximizes probability of training data subject to regularization which encourages low or sparse weights

LOGISTIC REGRESSION

- Pros: A simple yet low-bias classifier; unlike Naïve Bayes not confounded by diverse, correlated features
- Cons: Slow to train; some feature scaling issues; often needs a lot of data to work well; choosing regularisation a nuisance but important since overfitting is a big problem

K-NEAREST NEIGHBOUR

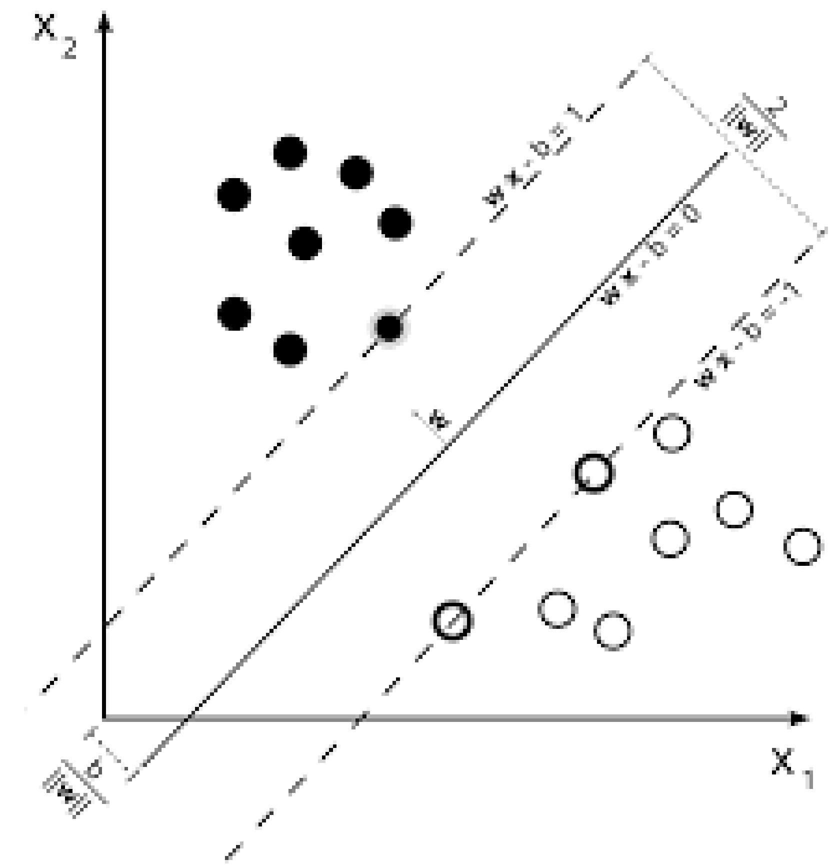
- Classify based on majority class of k-nearest training examples in feature space



- Definition of nearest can vary
  - Euclidean distance
  - Cosine distance
- Pros: Simple, effective; no training required; inherently multiclass; optimal with infinite data
- Cons: Have to select k; issues with unbalanced classes; often slow (need to find those k-neighbours); features must be selected carefully

SUPPORT VECTOR MACHINES

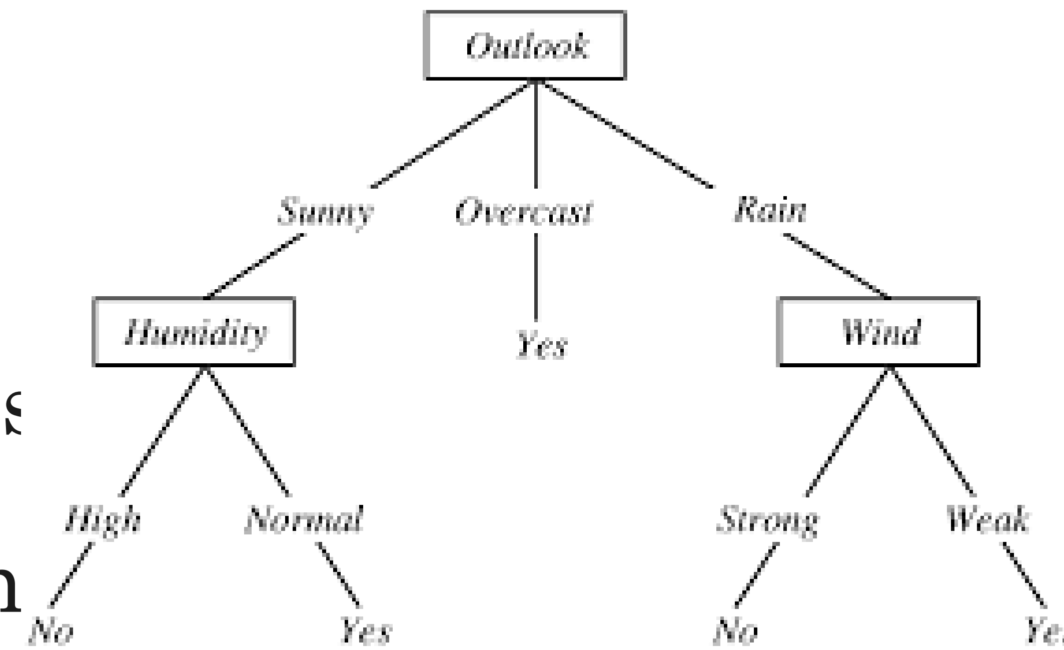
- Finds hyperplane which separates the training data with maximum margin
  - Allows for some misclassification
- Weight vector is a sum of support vectors (examples on the margin)



- Pros: fast and accurate linear classifier; can do non-linearity with kernel trick; works well with huge feature sets
- Cons: Multiclass classification awkward; feature scaling can be tricky; deals poorly with class imbalances; uninterpretable

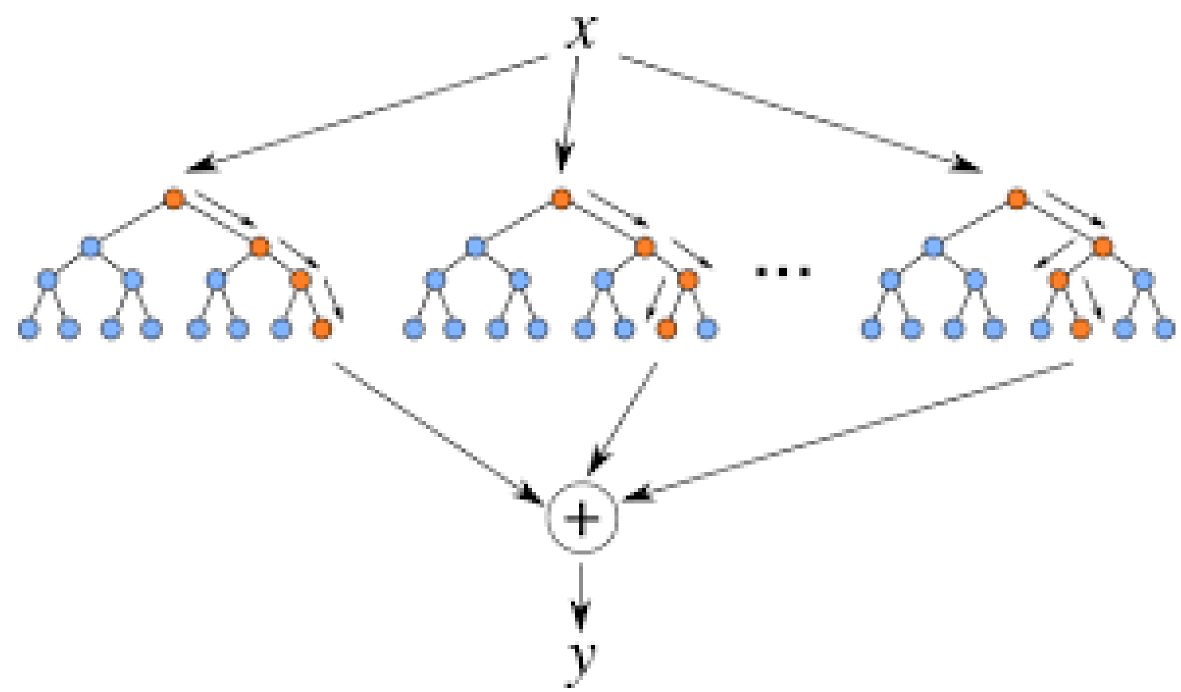
DECISION TREE

- Construct a tree where nodes correspond to tests on individual features



- Leaves are final class decisions
- Based on greedy maximization of mutual information
- Pros: in theory, very interpretable; fast to build and test; feature representation/scaling irrelevant; good for small feature sets, handles non-linearly-separable problems
- Cons: In practice, often not that interpretable; highly redundant sub-trees; not competitive for large feature sets

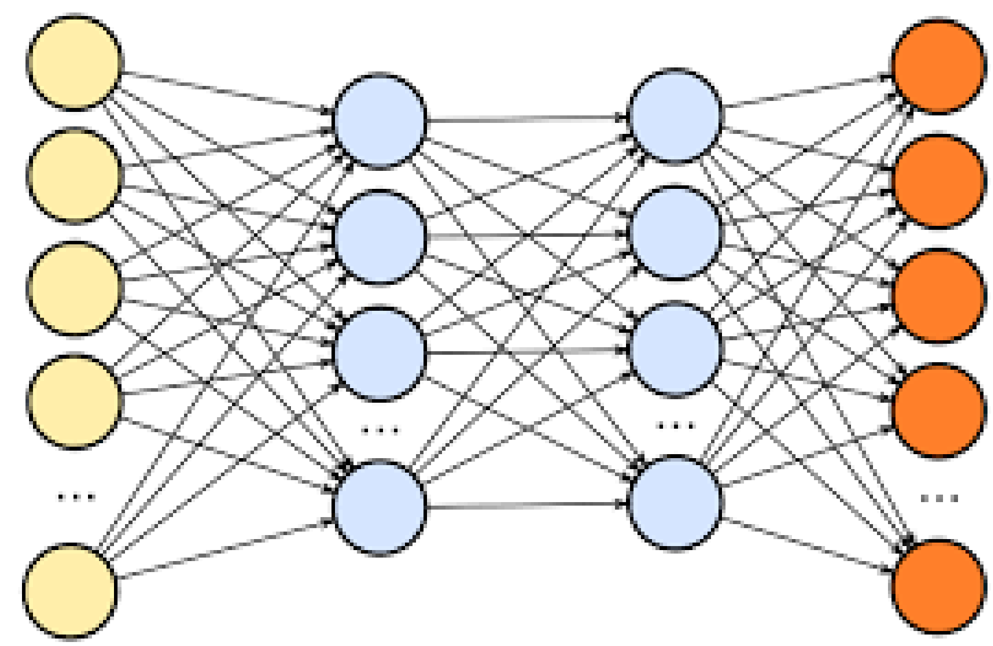
RANDOM FORESTS



- An *ensemble* classifier
- Consists of decision trees trained on different subsets of the training and feature space
- Final class decision is majority vote of sub-classifiers
- Pros: Usually more accurate and more robust than decision trees, a great classifier for small- to moderate-sized feature sets; training easily parallelised
- Cons: Same negatives as decision trees: too slow with large feature sets

NEURAL NETWORKS

- An interconnected set of nodes typically arranged in layers
- Input layer (features), output layer (class probabilities), and one or more hidden layers



- Each node performs a linear weighting of its inputs from previous layer, passes result through activation function to nodes in next layer
- Pros: Extremely powerful, state-of-the-art accuracy on many tasks in natural language processing and vision
- Cons: Not an off-the-shelf classifier, very difficult to choose good parameters; slow to train; prone to overfitting

TEXT CLASSIFICATION TASKS

- Obviously more than can be enumerated here
- But let's briefly look at some major examples:
  - Topic classification
  - Polarity classification
  - Genre classification
  - Authorship attribution
  - Native-language identification
- Not necessarily a full-sized "text"
  - E.g. sentence or tweet-level polarity classification

TOPIC CLASSIFICATION

- Motivation: library science, information retrieval
- Classes: Topic categories, e.g. "jobs", "anxiety disorders"
- Features
  - *Unigram* bag of words (BOW), with stop-words removed
  - Longer *n-grams (bigrams, trigrams*) for phrases
- Examples of corpora
  - Reuters news corpus (RCV1, see NLTK sample)
  - Pubmed abstracts
  - Tweets with hashtags

TOPIC CLASSIFICATION EXAMPLE

Is the topic of this text from the Reuters news corpus acquisitions or earnings?

LIEBERT CORP APPROVES MERGER

Liebert Corp said its shareholders approved the merger of a wholly-owned subsidiary of Emerson Electric Co. Under the terms of the merger, each Liebert shareholder will receive .3322 shares of Emerson stock for each Liebert share.

ANSWER: ACQUISITIONS

POLARITY CLASSIFICATION

- Motivation: Sentiment analysis, opinion mining
- Classes: Positive/Negative/(Neutral)
- Features
  - *N*-grams
  - Polarity lexicons
- Examples of corpora
  - Polarity movie review dataset (in NLTK)
  - SEMEVAL Twitter polarity datasets

POLARITY CLASSIFICATION EXAMPLE

What is the polarity of this tweet from the SEMEVAL dataset?

anyone having problems with Windows 10? may be coincidental but since i downloaded, my WiFi keeps dropping out. Itunes had a malfunction

ANSWER: NEGATIVE

GENRE CLASSIFICATION

- Motivation: Information retrieval, study of genre differences
- Classes: types of document, e.g. newswire, fiction, e-mails
- Features
  - Word N-grams
  - Structural features (spacing, punctuation)
  - Text statistics (word/sentence length, lexical density, frequency of word categories)
- Examples of corpora
  - Brown corpus (see NLTK sample)
  - British National Corpus (BNC)

GENRE CLASSIFICATION EXAMPLE

What is the genre of this text from the Brown corpus?

The expense and time involved are astronomical. However, we sent a third vessel out, a much smaller and faster one than the first two. We have learned much about interstellar drives since a hundred years ago; that is all I can tell you about them.

`"But the third ship came back several years ago and reported. That it had found a planet on which human beings could live and which was already inhabited by sentient beings!" said Hal, forgetting in his enthusiasm that he had not been asked to speak .

ANSWER: SCIENCE FICTION

AUTHORSHIP ATTRIBUTION

- Motivation: forensic linguistics, plagiarism detection
- Classes: Authors (e.g. Shakespeare)
- Features
  - Frequency of function words
  - Character n-grams
  - Discourse structure
- Examples of corpora
  - Project Gutenberg corpus (see NLTK sample)
  - Livejournal blog corpus

AUTHOR ATTRIBUTION EXAMPLE

Which famous novelist wrote this text from Project Gutenberg?

Mr. Dashwood's disappointment was, at first, severe; but his temper was cheerful and sanguine; and he might reasonably hope to live many years, and by living economically, lay by a considerable sum from the produce of an estate already large, and capable of almost immediate improvement. But the fortune, which had been so tardy in coming, was his only one twelvemonth. He survived his uncle no longer; and ten thousand pounds, including the late legacies, was all that remained for his widow and daughters.

ANSWER:JANE AUSTEN

NATIVE-LANGUAGE IDENTIFICATION

- Motivation: forensic linguistics, educational applications
- Classes: first language of author (e.g. Chinese)
- Features
  - Word N-grams
  - Syntactic patterns (POS, parse trees)
  - Phonological features
- Examples of corpora
  - TOEFL/IELTS essay corpora
  - Lang-8 language learner website

NATIVE-LANGUAGE IDENTIFICATION

What is the native language of the writer of this text?

Now a festival of my university is being held, and my club is joining it by offering a target practice game using bows and arrows of archery. I'm a manager of the attraction, so I have worked to make it succeed. I found it puzzled to manage a event or a program efficiently without generating a free rider.The event is not free, so we earn a lot of money.

ANSWER:JAPANESE

A FINAL WORD

- Lots of algorithms available to try out on your task of interest (see scikit-learn)
- But if good results on a new task are your goal, then well-annotated, plentiful datasets and appropriate features often more important than the specific algorithm used

FURTHER READING

- J&M3 Ch. 6,7
- *Optional*: For more in-depth discussion of machine-learning algorithms mentioned here (and more), *Pattern Recognition and Machine Learning* by Bishop is a good general introduction.