Information Retrieval (IN4325)

Evaluation (cont)+
Machine learning for NLP

Dr. Nava Tintarev
Assistant Professor, TU Delft

TU Delft

Last week

- Extrinsic versus intrinsic evaluation
- BLEU and Rouge
- Skip-gram
- Why not always use offline metrics?

Experimental Design

- **Hypotheses**
- **Subjects**
- **Material**
- Procedure...
- Analysis

TU Delft

Procedure: Ethics

- Can doing experiment harm people?
  - » BT-Nurse and patient care
  - » If so, must present acceptable solution
- Subjects can drop out at any time
  - » Can NOT “pressure” them to stay if the want to quit experiment
- Consent forms and ethics committee!

Procedure: Exclusion

- **When do we drop a subject from the experiment?**
  - » Incomplete responses?
  - » Inconsistent responses?
  - » Bizarre responses?
- **Human-computation**
  - » Acceptance rates
  - » Control questions
  - » Durations

TU Delft

Procedure: SumTime

- Questions
  - » Presented 2 variants
  - » Which variant is: easiest to read; most accurate; most appropriate
- Order not randomised
- No payment
- No practice or filler, no ethical issues
- Excluded if less than 50% completed

Statistics: Test

- Principle: Likert scales are not numbers
  - » Should not be averaged
  - » Non-parametric test (Wilcoxon Signed Rank)
- Practice
  - » Often present average Likert score
  - » Use parametric test, such as t-test
  - » More or less works….
    - – But not if rigorous stats needed!
    - – Need to check if data is normally distributed.

TU Delft

Statistics: Normalisation

- Some users are more generous than others
- Some scenarios are harder than others
- Potential bias
  - » User X always rates “Great”, Y always “Poor”
  - » X rates 10 SumTime texts and 1 corpus text
  - » Y rates 1 SumTime text and 10 corpus texts
- Use balanced design (Latin square)
- Use linear model
  - » Predicts score on user, scenario, presentation
  - » Just look at presentation element

Statistics: Multiple Hypoth

- Bonferroni multiple hypothesis correction
- Divide significance p value by number of hypotheses being tested
  - » 1 hypothesis: look for $p < .05$
  - » 2 hypotheses: look for $p < 0.025$
  - » 10 hypotheses: look for $p > 0.005$

TU Delft

Statistics: SumTime

- Test: Chi-square
  - » Because users asked to state a preference between variants, did not give Likert score
- Normalisation: not necessary
  - » Less important with preferences
    - – If user is asked whether A or B is better, does not matter how generous he is (“great” vs “poor”)
- Multiple hypotheses: $p < 0.025$
  - » Because 2 hypotheses

TU Delft

Which technique to use?

- Most common is laboratory ratings
  - But we know these may not correlate with task performance (e.g. Babytalk experiment)
- Task-based and/or real-world evaluation is harder, but more meaningful.
- Metrics should not be only evaluation
- Good experimental design and statistics!

TU Delft

A note on “null results"

- Method is important
- Critical analysis is important
  - Where and why do we have poor performance
- Comparison is important
- High accuracy/precision/recall is secondary!

This week

Machine learning for NLP

- Classes of machine learning problems
- Feature selection/extraction
- Common ML techniques
  - Discriminative: SVM, MaxEnt
  - Generative: NB, logistic regression

Credits

Stanford NLP course

Zoltán Szlávik IBM

- http://web.stanford.edu/class/cs276/handouts/lecture14-learning-ranking.ppt
- https://sites.google.com/a/unal.edu.co/information-retrieval-2015-1/
- http://mklab.iti.gr/essir2015/wp-content/uploads/2015/03/ESSIR2015_Sebastiani.pdf
- https://www.cs.utexas.edu/~mooney/cs391L/slides/svm.ppt
- http://web.stanford.edu/class/cs276/handouts/lecture12-clustering.ppt

Is this spam?

ACM TechNews, Friday, March 9, 2018

AT

ACM TechNews <technews-editor@acm.org>
Friday, 9 March 2018 at 18:52
To: Nava Tintarev - EWI

Unsubscribe Manage Add-ins...

Click here to view this online

acm TechNews

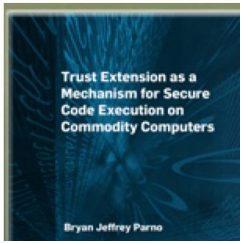
Computer science, information technology, and related science, society, and technology news

MEMBERNET ARCHIVES CAREERNEWS CONTACT US

Welcome to the March 9, 2018 edition of ACM TechNews, providing timely information for IT professionals three times a week.

ACM TechNews mobile apps are available for Android phones and tablets (click here) and for iPhones (click here) and iPads (click here).

To view "Headlines At A Glance," hit the link labeled "Click here to view this online" found at the top of the page in the html version. The online version now has a button at the top labeled "Show Headlines."



Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

TU Delft

Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

What is the subject of this article?

MEDLINE Article


Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …

Types of machine learning problems

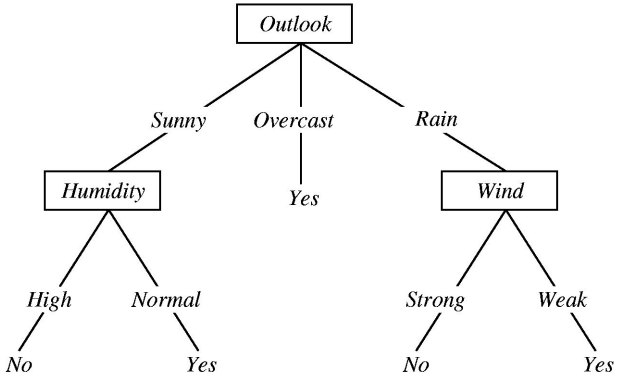
What is classification?

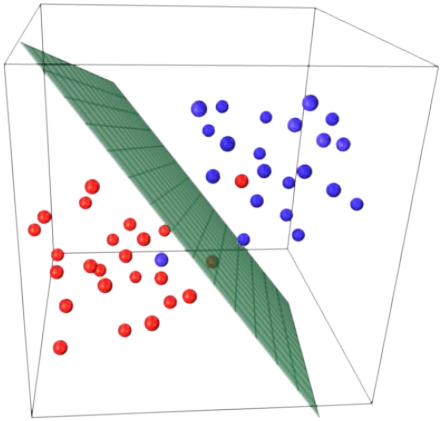
- Classification (aka "categorization"): a ubiquitous enabling technology in data science; studied within pattern recognition, statistics, and machine learning.

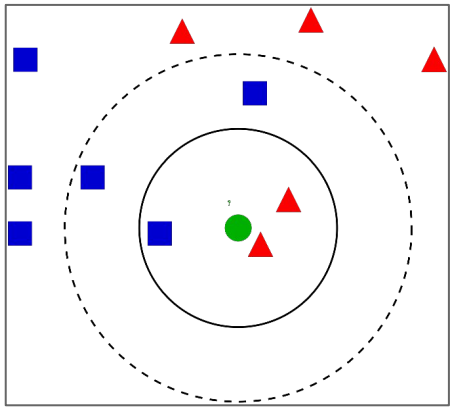
- Def: the activity of predicting to which among a predefined finite set of groups ("classes", or "categories", or "labels") a data item belongs to

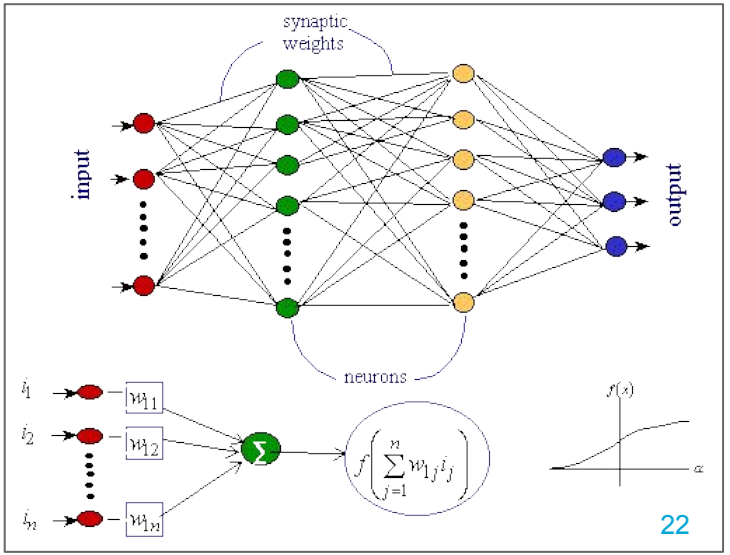
- Formulated as the task of generating a hypothesis (or "classifier", or "model") $h : D \rightarrow C$, where $D = \{\mathbf{x}_1, \mathbf{x}_2, ...\}$ is a domain of data items and $C = \{c_1, ..., c_n\}$ is a finite set of classes (the classification scheme)

Classifier examples









What is classification?

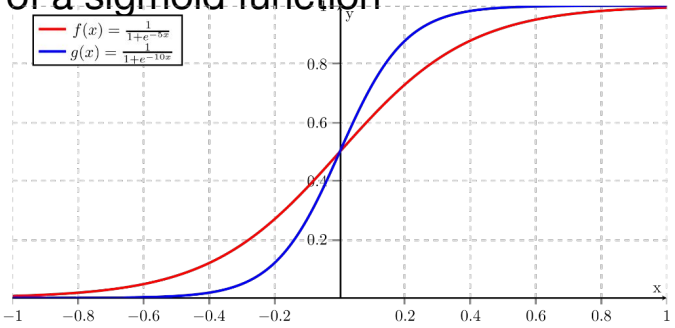
- Different from most clustering, where the groups ("clusters") and their number are not known in advance.
- In text classification, data items are textual (e.g., news articles, emails, sentences, queries, etc.) or partly textual (e.g., Web pages)
- The membership of a data item into a class must not be determinable with certainty (e.g., predicting whether a natural number belongs to Prime or NonPrime is not classification problem)

Types of classification

- Binary classification: each item belongs to exactly one class
  - E.g., assigning emails to Spam or Legitimate
- Single-Label Multi-Class (SLMC) classification: each item belongs to exactly one class
  - E.g., assigning news articles to one of HomeNews, International, Entertainment, Lifestyles, Sports
- Multi-Label Multi-Class (MLMC) classification: each item may belong to zero, one, or several classes)
  - E.g., assigning CS articles to classes in the ACM Classification System
  - May be solved as *n* independent binary classification problems
- Ordinal classification (OC): as in SLMC, but for the fact that there is an ordering of the class labels
  - E.g., assigning product reviews to one of Excellent, Good, SoAndSo, Poor, Disastrous

Hard vs. Soft Classification

- The definitions above denote "hard classification" (HC)
- "Soft classification" (SC) denotes the task of predicting a score for each pair (*d, c*), where the score denotes the { probability / strength of evidence / confidence } that *d* belongs to *c*
  - E.g., a probabilistic classifier outputs "posterior probabilities"
  - E.g., some classifiers output scores that represent their confidence that *d* belongs to *c* (values might fall outside of the 0-1 range)
- When scores are not probabilities, they can be converted into probabilities via the use of a sigmoid function



25

SC to HC

Hard classification often consists of

1. Training a soft classifier that outputs scores $s(d, c)$
2. Picking a threshold t, such that
   - $s(d, c) > t$ is interpreted as a "Yes"
   - Otherwise it's interpreted as a "No"

- In soft classification, scores are used for ranking; e.g., ranking items for a given class, ranking classes for a given item.

- HC is used for fully independent (no humans) classifiers, while SC is used for interactive classifiers (i.e., with humans in the loop).

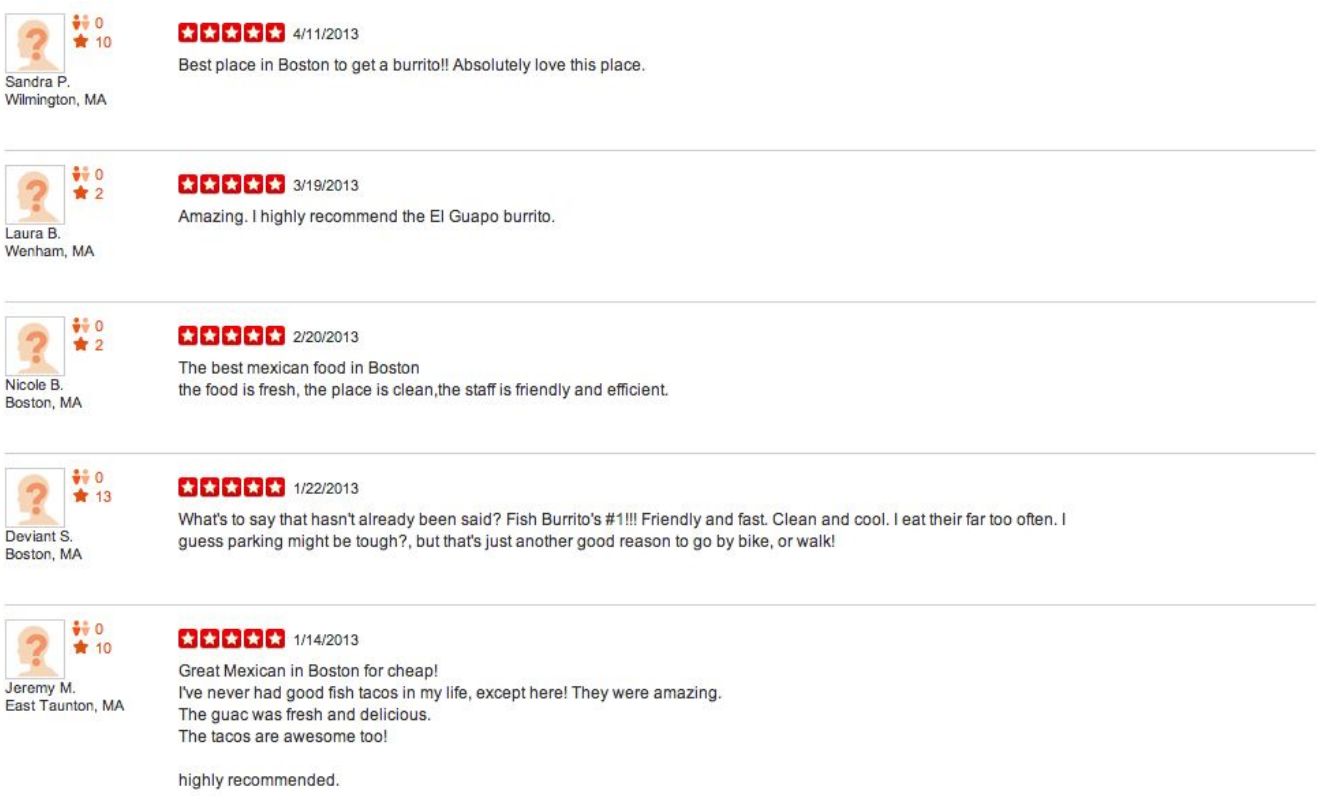
Dimensions of classification

- Text classification may be performed according to several dimensions (“axes”) orthogonal to each other
  - by topic; by far the most frequent case, its applications are ubiquitous
  - by sentiment; useful in market research, online reputation management, social science and political science
  - by veracity; e.g., fake news or reviews;
  - by author (aka “authorship attribution”), by native language (“native language identification”), or by gender ; useful in forensics and cybersecurity
  - by usefulness; e.g., product reviews
  - …

27

Fake review detection

How would you do it?



0
10
Sandra P.
Wilmington, MA

4/11/2013

Best place in Boston to get a burrito!! Absolutely love this place.

0
2
Laura B.
Wenham, MA

3/19/2013

Amazing. I highly recommend the El Guapo burrito.

0
2
Nicole B.
Boston, MA

2/20/2013

The best mexican food in Boston
the food is fresh, the place is clean,the staff is friendly and efficient.

0
13
Deviant S.
Boston, MA

1/22/2013

What's to say that hasn't already been said? Fish Burrito's #1!!! Friendly and fast. Clean and cool. I eat their far too often. I guess parking might be tough?, but that's just another good reason to go by bike, or walk!

0
10
Jeremy M.
East Taunton, MA

1/14/2013

Great Mexican in Boston for cheap!
I've never had good fish tacos in my life, except here! They were amazing.
The guac was fresh and delicious.
The tacos are awesome too!

highly recommended.

Your data

400 fake reviews from AMT and 400 non-fake reviews from Tripadvisor

Filtered and unfiltered Yelp reviews across 85 hotels and 130 restaurants (unbalanced)

Feature selection

Feature selection

Features	# of features	frequency or presence?
unigrams	16165	freq.
unigrams	”	pres.
unigrams+bigrams	32330	pres.
bigrams	16165	pres.
unigrams+POS	16695	pres.
adjectives	2633	pres.
top 2633 unigrams	2633	pres.
unigrams+position	22430	pres.

- Unigrams, bigrams
- Frequency or presence?
- POS, (only) adjectives
- Position: first quarter, last quarter, or middle half of the document

Content

- In order to be input to a learning algorithm (or a classifier), all training (or unlabeled) documents are converted into vectors in a common vector space
- The dimensions of the vector space are called features
- In order to generate a vector-based representation for a set of documents *D*, the following steps need to be taken
    1. Feature Extraction (+ Weighting)
    2. (Feature Selection *or* Feature Synthesis)
    3. Feature Weighting

Feature extraction

- For instance, for classification by topic you might want to have features representing words
- Other classification dimensions, however, might require different kinds of features
- The choice of features for a classification task (feature design) is dictated by the distinctions we want to capture, and is left to the designer; e.g.
  - in classification by author, features such average word length, average sentence length, punctuation frequency, frequency of subjunctive clauses, etc., are used
  - in classification by sentiment, bag-of-words is not enough, and deeper linguistic processing is necessary

Feature synthesis

- Matrix decomposition techniques (e.g., PCA, SVD, LSA) can be used to synthesize new features that replace the features discussed above
- These techniques are based on the principles of distributional semantics, which states that the semantics of a word “is” the words it co-occurs with in corpora of language use
- The advantage of these techniques is that the synthetic features in the new vectorial representation do not suffer from problems such as polysemy and synonymy
- The disadvantage of these techniques is that they are computationally expensive, sometimes prohibitively so

Feature weighting

- Feature weighting means attributing a value to a feature in a document: this value may be
  - binary (representing presence/absence of the word in the document); or
  - numeric (representing the importance of the word in the document); obtained via feature weighting functions in the following two classes:
    - unsupervised: e.g, *tfidf* or *BM25*,
    - supervised: e.g., *tf*IG*, *TF*chi*^{2}

Feature selection

- Vectors of length $O(10^5)$ or $O(10^6)$ may result, esp. if word *n*-grams are used; this may give rise to both overfitting and high computational cost;
- Feature selection (FS) has the goal of identifying the most discriminative features, so that the others may be discarded
- The “filter” approach to FS consists in measuring (via a function) the discriminative power of each feature and retaining only the top-scoring features
- Typical measures:
  - Mutual information, chi-square, log-odds

Enter Deep Learning – do we really need to engineer all those features?

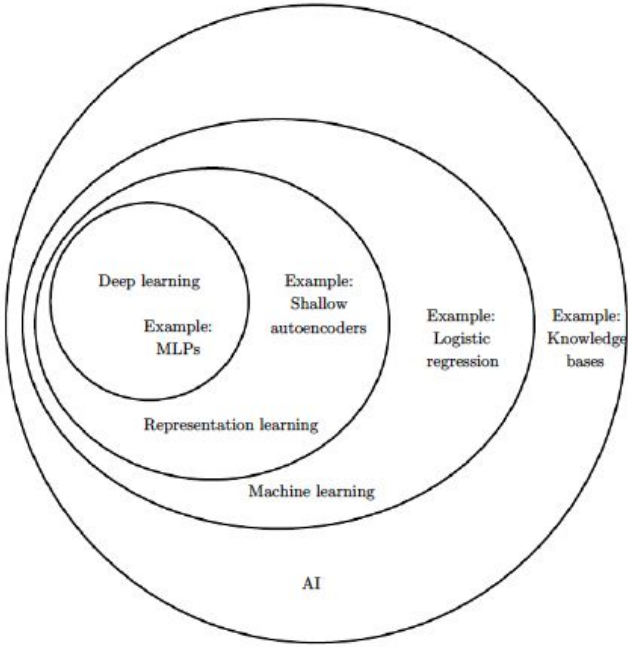
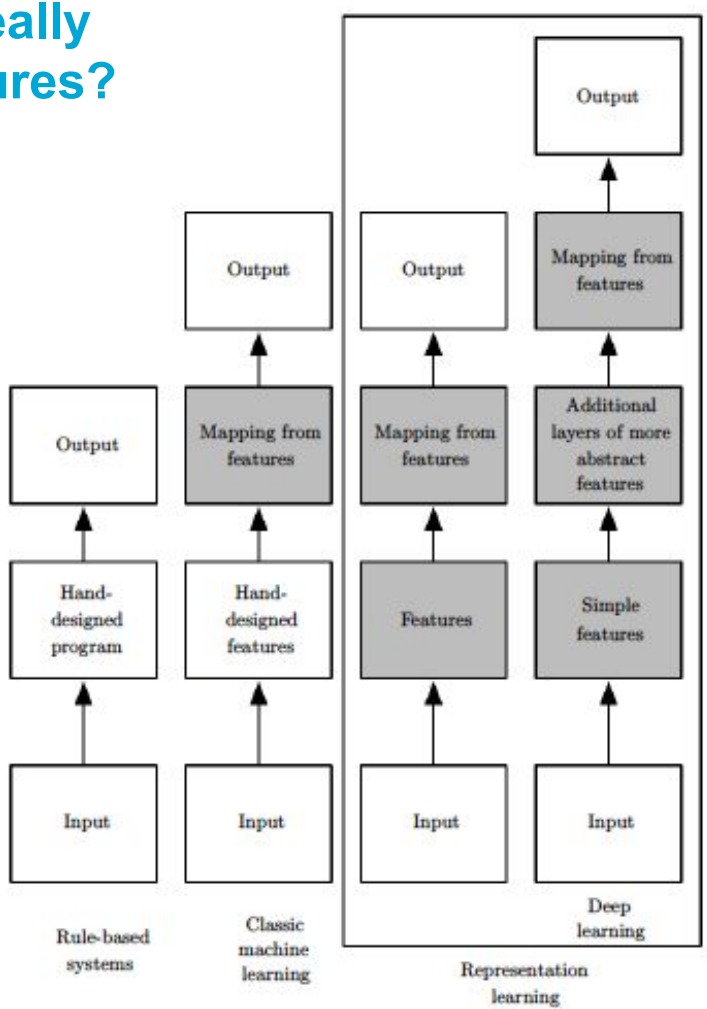
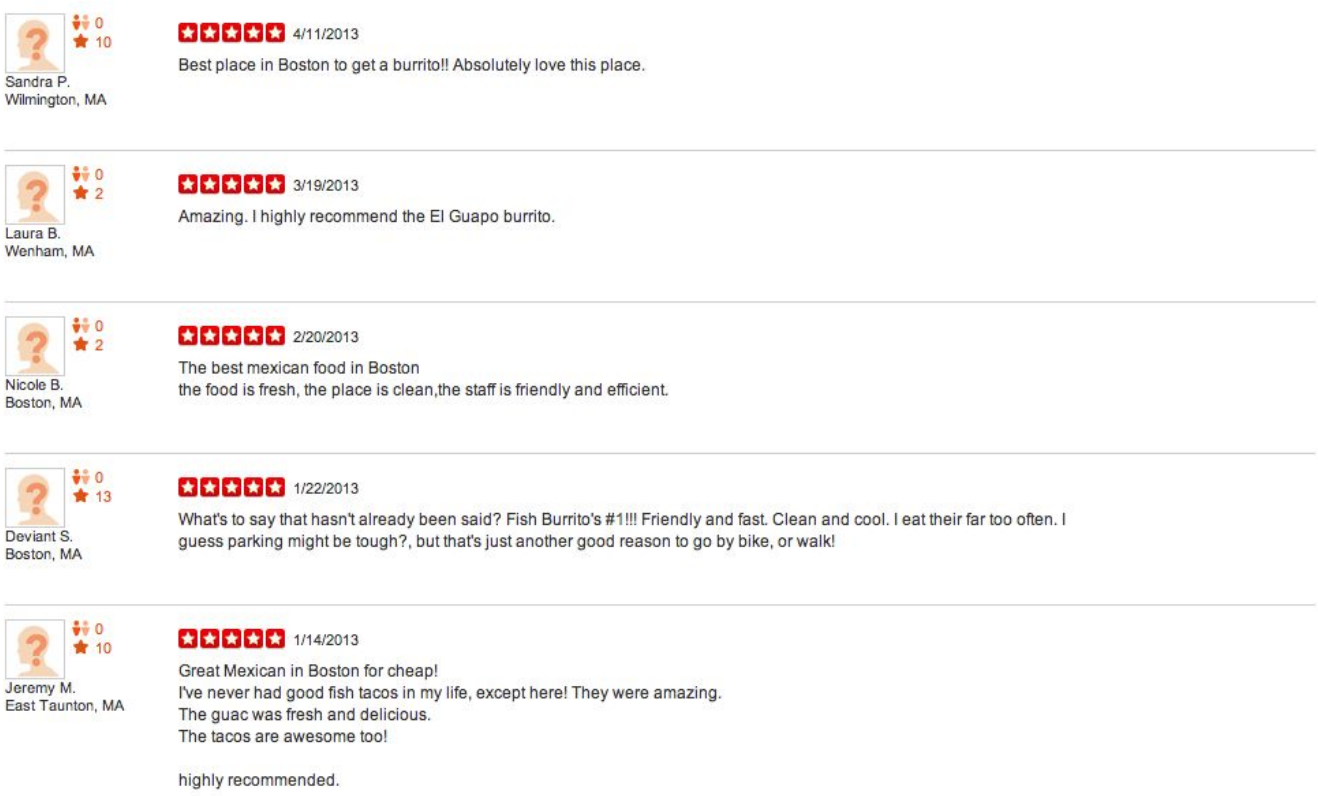


Figure 1.4: A Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Each section of the Venn diagram includes an example of an AI technology.



Fake review detection

How would you do it?



0
10
Sandra P.
Wilmington, MA

4/11/2013

Best place in Boston to get a burrito!! Absolutely love this place.

0
2
Laura B.
Wenham, MA

3/19/2013

Amazing. I highly recommend the El Guapo burrito.

0
2
Nicole B.
Boston, MA

2/20/2013

The best mexican food in Boston
the food is fresh, the place is clean,the staff is friendly and efficient.

0
13
Deviant S.
Boston, MA

1/22/2013

What's to say that hasn't already been said? Fish Burrito's #1!!! Friendly and fast. Clean and cool. I eat their far too often. I guess parking might be tough?, but that's just another good reason to go by bike, or walk!

0
10
Jeremy M.
East Taunton, MA

1/14/2013

Great Mexican in Boston for cheap!
I've never had good fish tacos in my life, except here! They were amazing.
The guac was fresh and delicious.
The tacos are awesome too!

highly recommended.

Your data

How would you do it?

400 fake reviews from AMT and 400 non-fake reviews from Tripadvisor

Filtered and unfiltered Yelp reviews across 85 hotels and 130 restaurants (unbalanced)

TU Delft

Common Machine Learning techniques

Machine learning techniques

	NB	ME	SVM
77.0	**77.7**	75.1
80.3	81.0	**81.4**
81.0	80.1	**81.6**

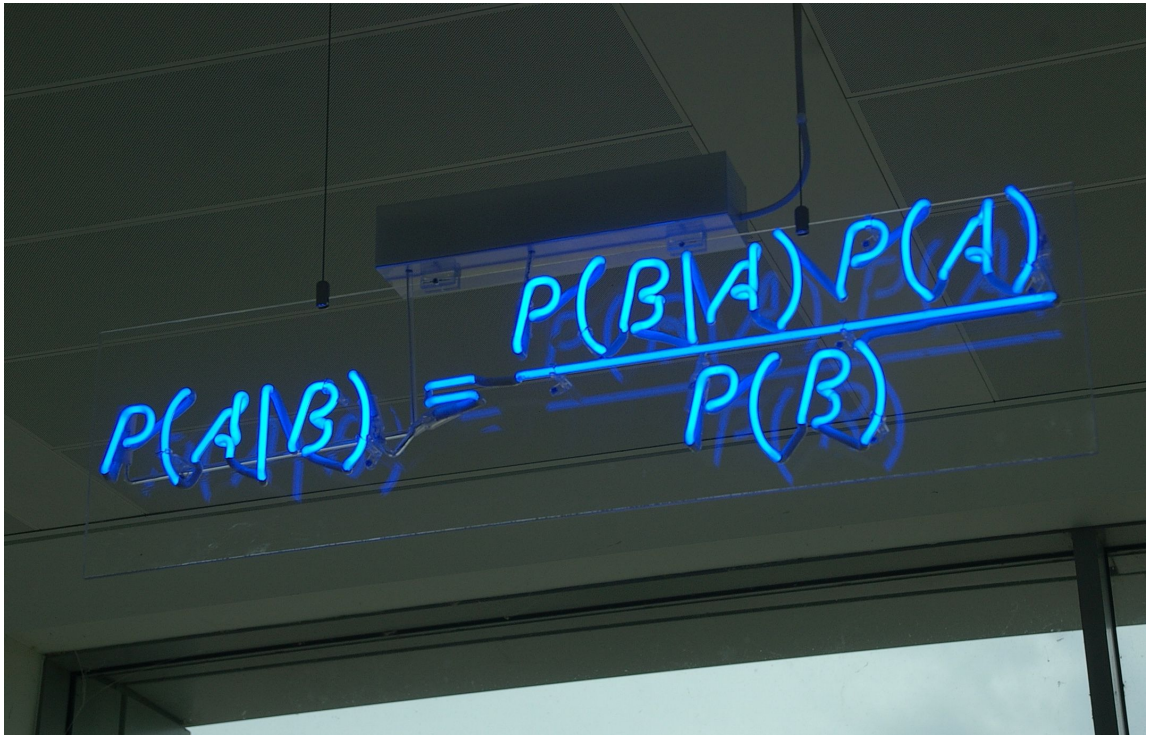
	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	”	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

- Naive-Bayes
- Maximal Entropy
- Support Vector Machine

Common Machine Learning techniques

- Common ML techniques
  - **Generative: NB**
  - Discriminative: SVM, MaxEnt/Log. Reg.
  - Discriminative v. Generative

Naive Bayes



Naïve Bayes

- Classify based on prior weight of class and conditional parameter for what each word says:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \left[\log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j) \right]$$

- Training is done by counting and dividing:

$$P(c_j) \leftarrow \frac{N_{c_j}}{N} \qquad P(x_k \mid c_j) \leftarrow \frac{T_{c_j x_k} + \alpha}{\sum_{x_i \in V} [T_{c_j x_i} + \alpha]}$$

- Don't forget to smooth

SpamAssassin

- Naïve Bayes has found a home in spam filtering
  - Widely used in spam filters
  - Many features beyond words

SpamAssassin Features:

- Basic (Naïve) Bayes spam probability
  - **Mentions:** Generic Viagra
  - **Regex:** millions of (dollar) ((dollar) NN,NNN,NNN.NN)
  - **Phrase:** impress ... girl
  - **Phrase:** 'Prestigious Non-Accredited Universities'
  - **From:** starts with many numbers
  - Subject is all capitals
  - HTML has a low ratio of text to image area
  - ...
  - https://spamassassin.apache.org/old/tests_3_3_x.html

Naive Bayes is Not So Naive

- Very fast learning and testing (basically just count features)
- Low storage requirements
- Very good in domains with many equally important features
- More robust to irrelevant features than many learning methods

  Irrelevant features cancel out without affecting results

Naive Bayes is Not So Naive

- More robust to concept drift (changing class definition over time)
- A good dependable baseline for text classification (but not the best)!

Naive Bayes vs. MaxEnt Models

- Naive Bayes models multi-count correlated evidence
  - Each feature is multiplied in, even when you have multiple features telling you the same thing
- Maximum Entropy models (pretty much) solve this problem
  - This is done by weighting features so that model expectations match the observed (empirical) expectations

Questions?

Common Machine Learning techniques

- Common ML techniques
  - Generative: NB
  - **Discriminative: SVM, MaxEnt/Log. Reg.**
  - Discriminative v. Generative

Support Vector Machines

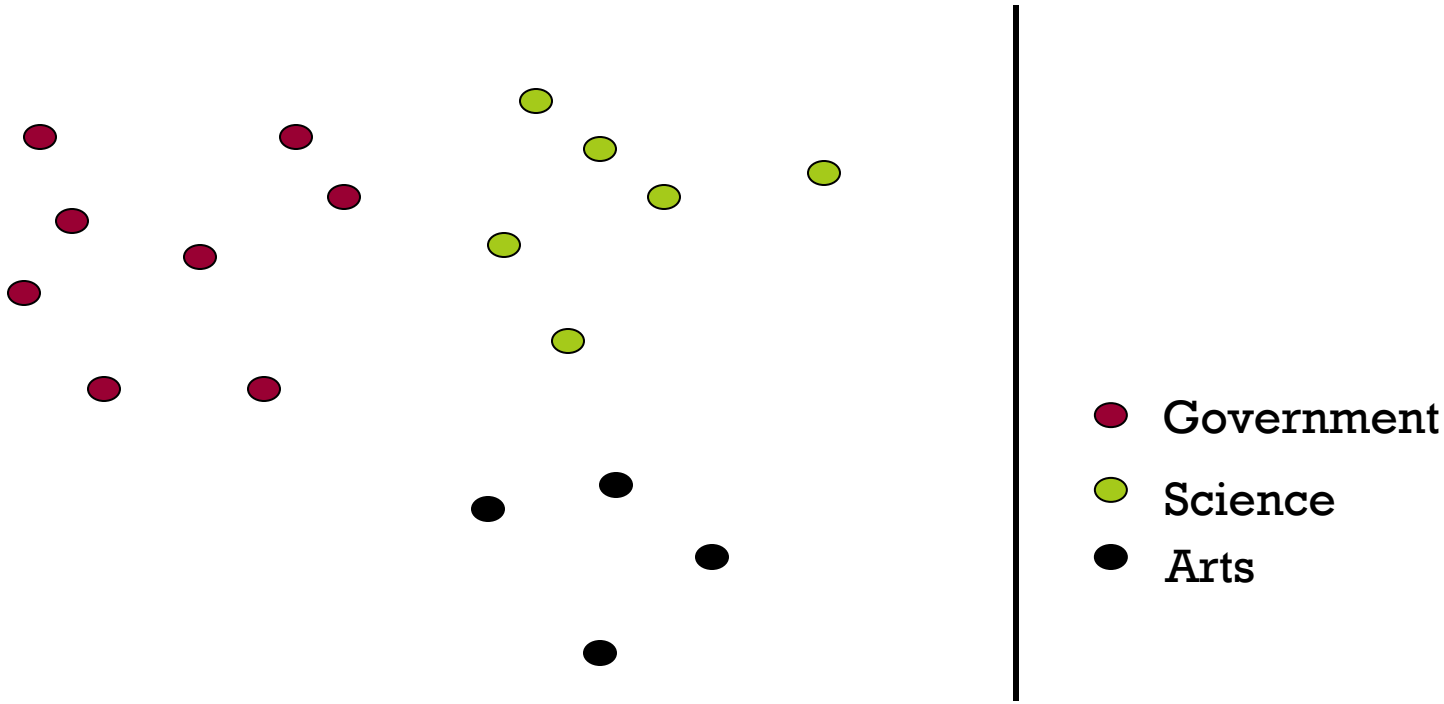
In vector space classification, training set corresponds to a labeled set of points (equivalently, vectors)

Premise 1: Documents in the same class form a contiguous region of space

Premise 2: Documents from different classes don't overlap (much)

Learning a classifier: build surfaces to delineate classes in the space

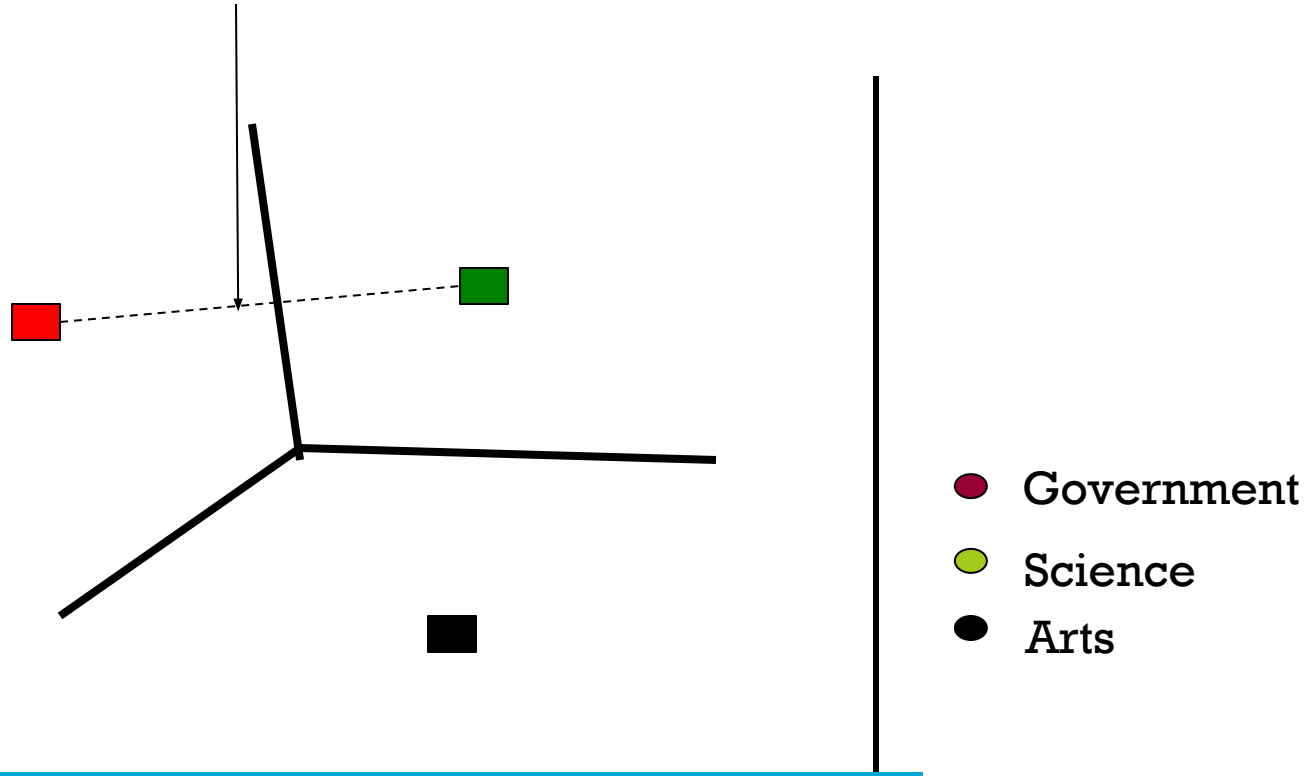
Documents in a Vector Space



Test Document, of what class?

- Government
- Science
- Arts

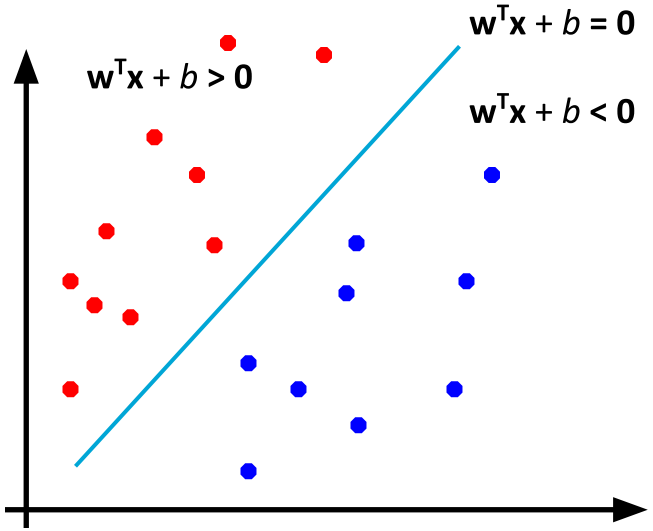
Test Document = Government



Our focus: how to find good separators

Linear

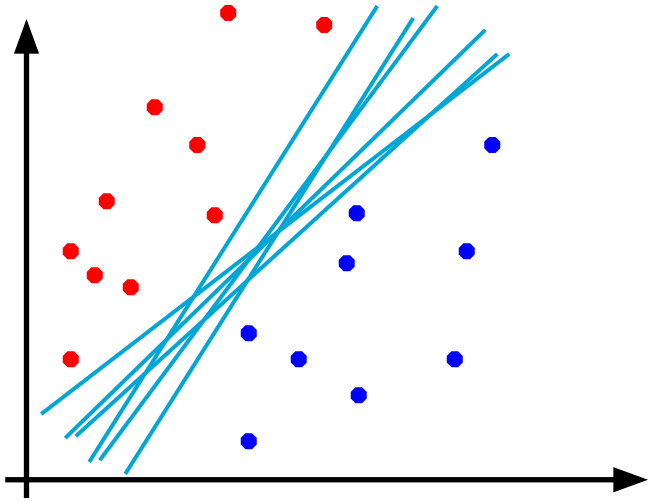
- Binary classification can be viewed as the task of separating classes in feature space:



$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$

Linear Separators

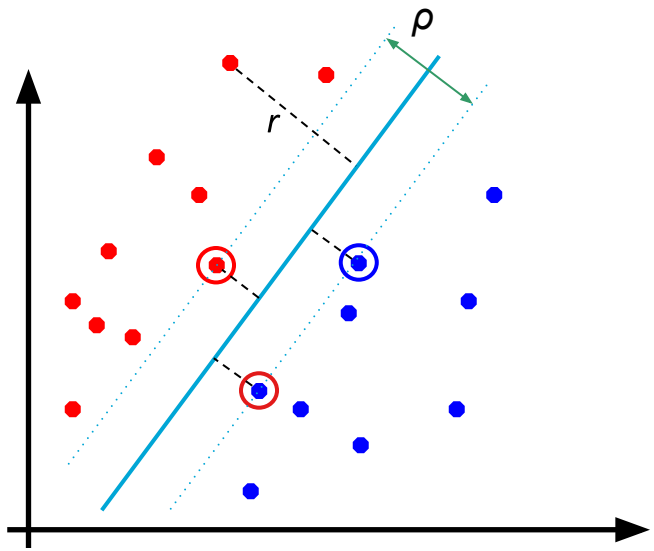
- Which of the linear separators is optimal?



57

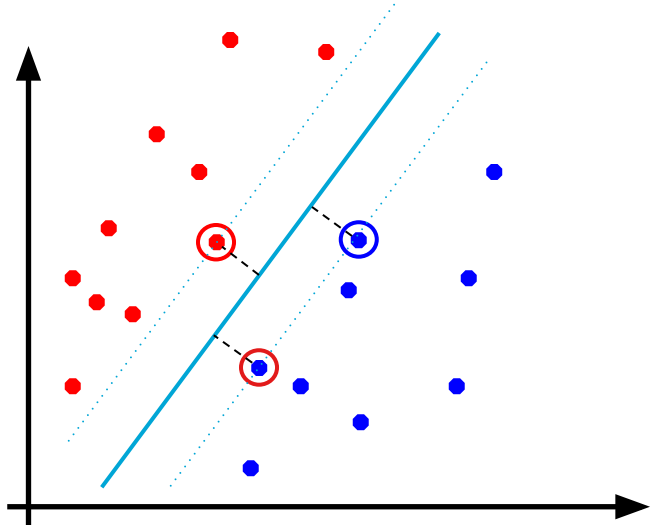
Classification Margin

- Compute distance from points to the separator
- Examples closest to the hyperplane are ***support vectors***.
- ***Margin*** ρ of the separator is the distance between support vectors.



Maximum Margin Classification

- Maximizing the margin is good according to intuition and PAC (**probably approximately correct learning)** theory.
- Implies that only support vectors matter; other training examples are ignorable.

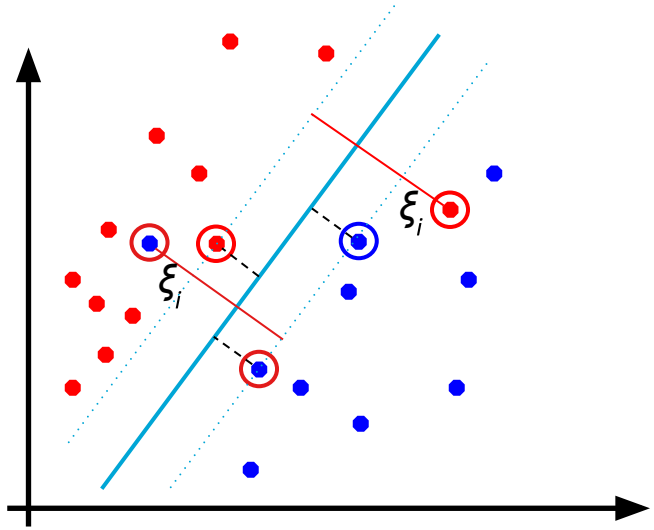


Optimization Problem

- We need to optimize a function subject to some (*linear)* constraints.
- Quadratic optimization problems are a well-known class of mathematical programming problems for which several (non-trivial) algorithms exist.

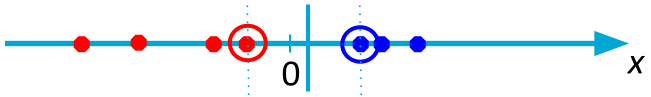
Soft Margin Classification

- What if the training set is not linearly separable?
- *Slack variables* ξ_i can be added to allow misclassification of difficult or noisy examples, resulting margin called *soft*.

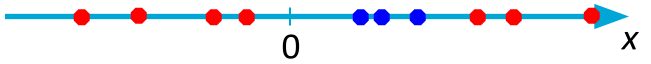


Non-linear SVMs

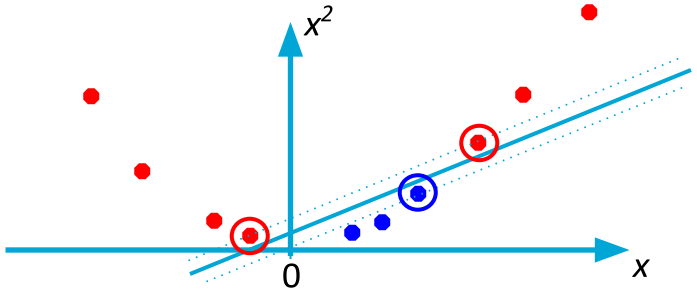
Datasets that are linearly separable with some noise work out great:



But what are we going to do if the dataset is just too hard?

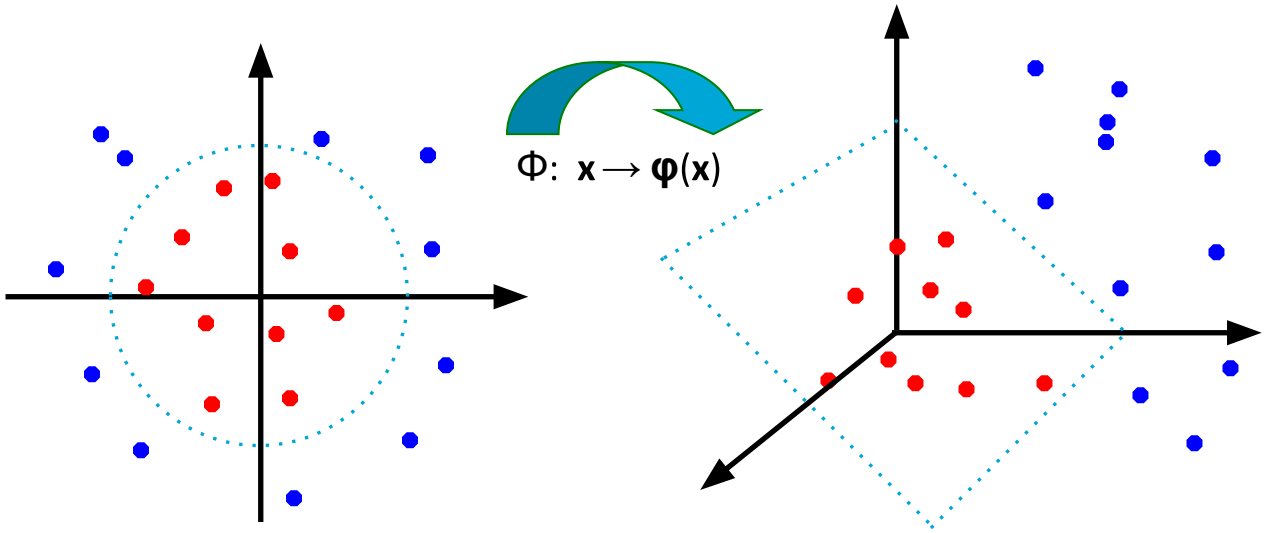


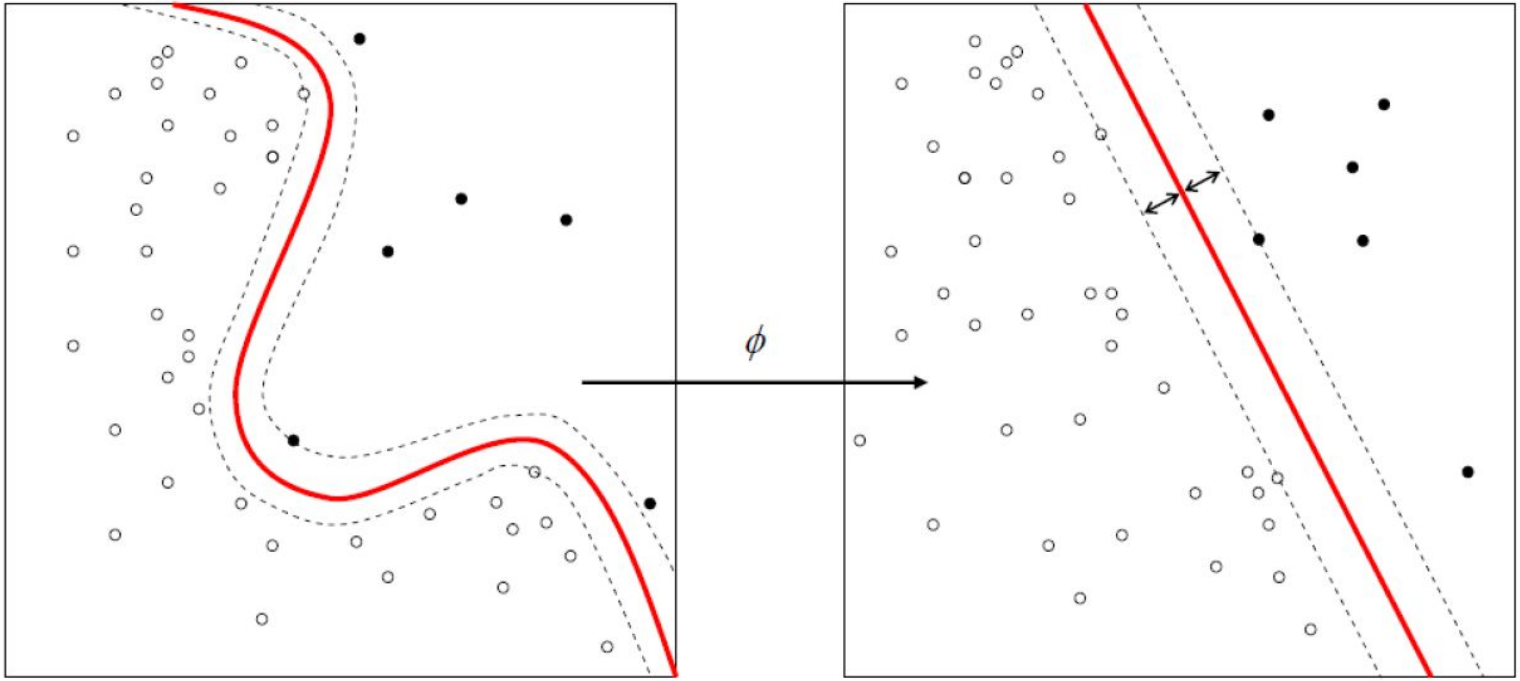
How about… mapping data to a higher-dimensional space:



Non-linear

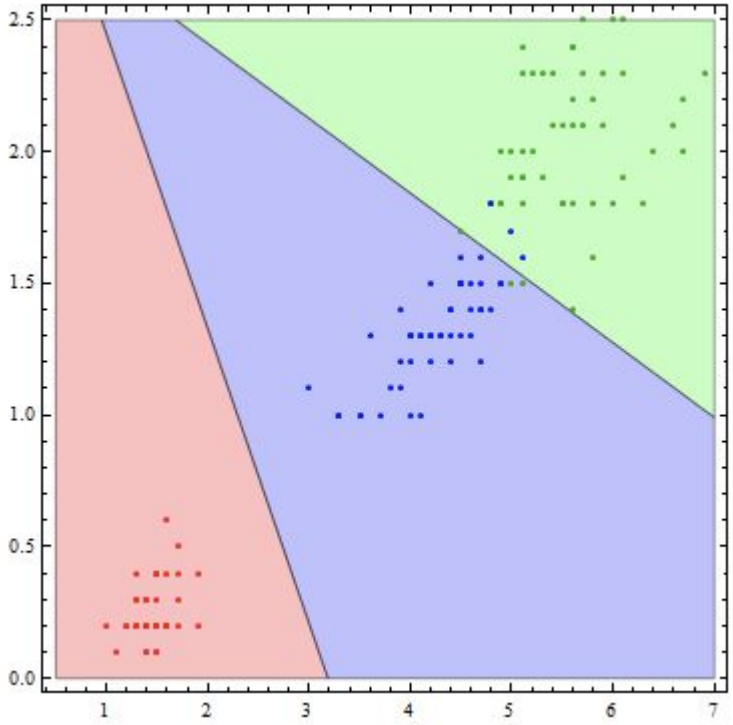
- General idea: the original feature space can be mapped to some higher-dimensional feature space where the training set is separable:





64

Maximum Entropy Classifiers



Maximum Likelihood Estimates

Maximizes the likelihood of the training set T given the model M

- What is the probability that a random word from some other text will be “banana”?
- Also called Maximum Entropy Classifiers
- Very similar to Naive Bayes
- Sets the models parameters using optimization
  - **Aim:** Maximize the performance of the classifier. “Set of parameters that maximizes the **total likelihood** of the training corpus”.
  - Uses iterative optimization techniques
    - Initialize and optimize

Exponential Model Likelihood

Maximum (Conditional) Likelihood Models :

- Given a model form, choose values of parameters to maximize the (conditional) likelihood of the data.

$$\log P(C \mid D, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c \mid d, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}$$

Maximum Likelihood Estimates

Comparison with Naive Bayes

Parameter can associate:

- a feature with more than one label;
- or more than one feature with a given label.

- In contrast, Naive-Bayes: one parameter per label (prior), and one parameter for (feature, label) pairs.

Maximum Likelihood Estimates

- Multinomial logistic regression
  - Extracting some set of weighted features from the input,
  - taking logarithms
  - combining them linearly

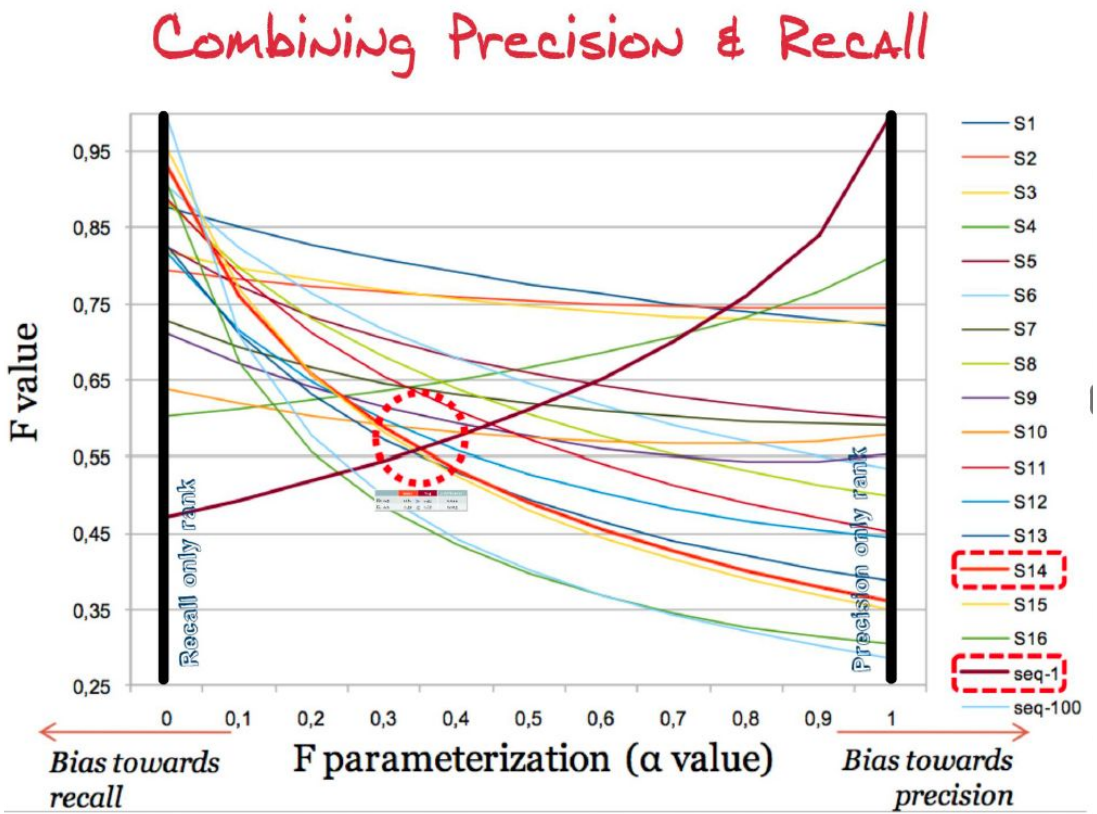
- **Logistic regression** - refers to a classifier that classifies an observation into one of two classes (e.g., Spam/noSpam)
- **Multinomial logistic regression** - classifying into more than two classes (Single label, multi-class)

Maximum Likelihood Estimates

- Given this model form, we will choose parameters $\{\lambda_i\}$ that *maximize the conditional likelihood* of the data according to this model.
  - We construct not only classifications, but probability distributions over classifications.
  - Gives a probability distribution over the classes (soft classification).

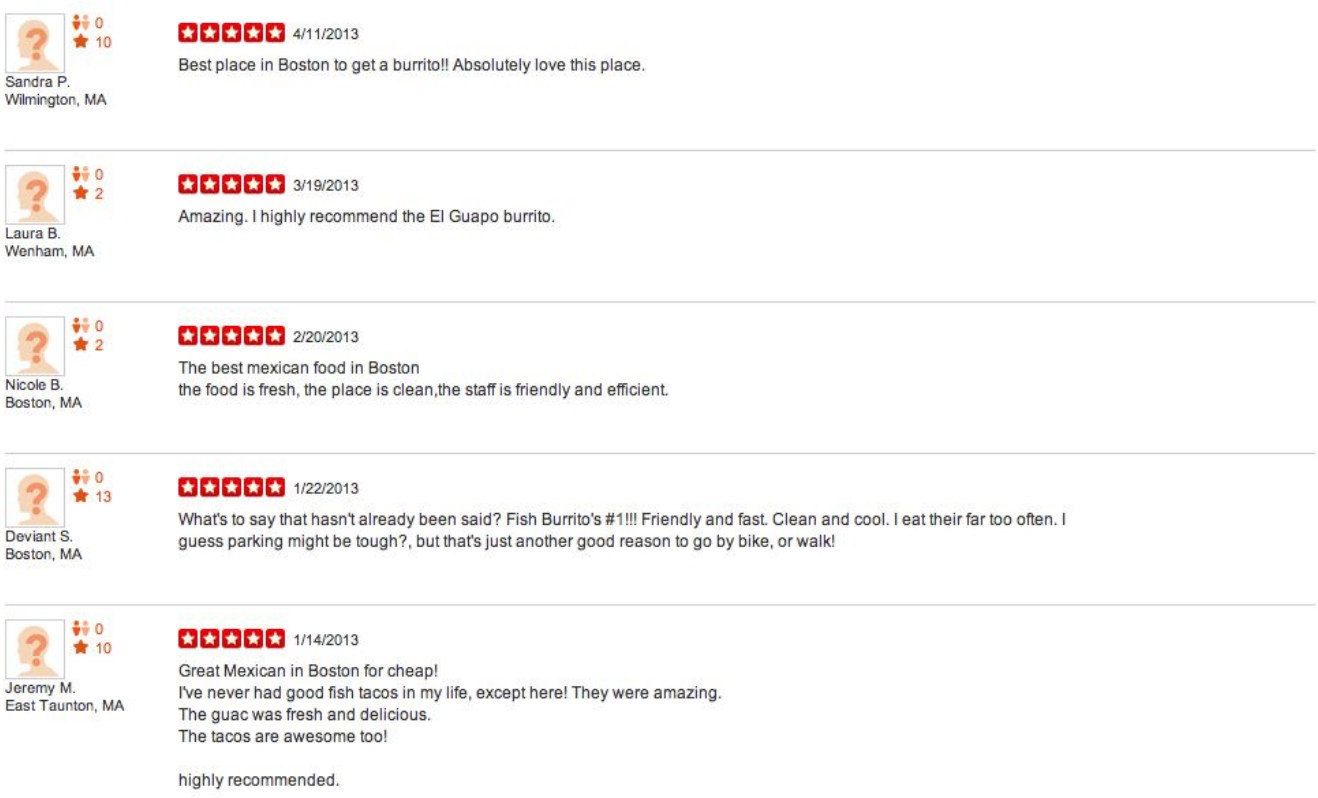
A note on F-scores

Amigó, Enrique, et al. "Combining evaluation metrics via the unanimous improvement ratio and its application to clustering tasks." *Journal of Artificial Intelligence Research* 42 (2011): 689-718



Fake review detection

How would you do it?



0
10
Sandra P.
Wilmington, MA

4/11/2013

Best place in Boston to get a burrito!! Absolutely love this place.

0
2
Laura B.
Wenham, MA

3/19/2013

Amazing. I highly recommend the El Guapo burrito.

0
2
Nicole B.
Boston, MA

2/20/2013

The best mexican food in Boston
the food is fresh, the place is clean,the staff is friendly and efficient.

0
13
Deviant S.
Boston, MA

1/22/2013

What's to say that hasn't already been said? Fish Burrito's #1!!! Friendly and fast. Clean and cool. I eat their far too often. I guess parking might be tough?, but that's just another good reason to go by bike, or walk!

0
10
Jeremy M.
East Taunton, MA

1/14/2013

Great Mexican in Boston for cheap!
I've never had good fish tacos in my life, except here! They were amazing.
The guac was fresh and delicious.
The tacos are awesome too!

highly recommended.

Mukherjee et al. What Yelp Fake Review Filter Might Be Doing? ACL-13

Fake Review Detection

SVM 5 fold cross-validation

Linguistic features

- Unigram, Bigram, PoS, Deep Syntax (production rules)

Behavioral Features

- Maximum number of reviews by a reviewer/day
- Percentage of positive reviews
- Review length
- Reviewer deviation (for business, then all his reviews)
- Maximum content similarity (for a given reviewer)

TU Delft

NLP intermediate project report

due (next) Wednesday March 4

- Title, authors
- Abstract
- **Introduction:** problem statement, motivation for the problem, overall plan to tackle the problem
- **Background:** what important works does this project build on
- **Approach:** what methods/algorithms did you use
- **Experiments:** describe your evaluation/experiments, the results and discuss them
- **Conclusions:** describe what you learnt/found and what avenues for future work you see
- References

This week

- Types of machine learning problems
- Feature selection/extraction
- Common ML techniques
  - Generative: NB
  - Discriminative: SVM, MaxEnt/Log. Reg.

Next lecture

- NLP ML continues
  - Discriminative versus Generative
  - Practical issues
- Natural language generation

Questions?