

# Information Retrieval (IN4325)

Machine learning for NLP

**Dr. Nava Tintarev**

**Assistant Professor, TU Delft**

# Last week

Bonferroni correction

Intrinsic vs extrinsic evaluation

Real world vs lab

Likert scale

BLEU vs ROUGE

# Experimental Design

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

# 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 they 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

# 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.

# 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  $p > 0.005$

# 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

# 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!

# Challenges

- Education and spread of best practice
- Design cheap/quick human evaluation that correlate with high-quality human evaluation
- Well-validated metrics

# Personal Lessons

- Publish negative results
  - Edge cases and outliers matter
  - Make sure code is debugged
  - Validate metrics
- 
- Not rocket science....still very important!
  - Do it properly and rigorously, with replicable results!
  - Otherwise it is not science

# Last week

## There are a lot of ways to evaluate...

- Task (extrinsic) evaluation
- Human ratings (intrinsic) evaluation
- Metric evaluation
- Setting up statistical tests
- Concluding thoughts

# This week

## Machine learning for NLP

- Classes of machine learning problems
- Feature selection/extraction
- Common ML techniques
  - Discriminative: SVM, MaxEnt
  - Generative: NB, logistic regression
  - Discriminative v. Generative
- Application domains
  - NER
  - Fake review detection

# 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!
- What have you learned?



## P9: Offspring from reproduction problems: What replication failure teaches us

Fokkens, Antske, et al. "Offspring from reproduction problems: What replication failure teaches us." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2013

# This week

- Intro
- Types of machine learning problems
- Feature selection/extraction
- Common ML techniques
  - Generative: NB
  - Discriminative: SVM, MaxEnt/Log.Reg.
  - Discriminative vs. Generative
- Application domain: Fake review detection

# Is this spam?

ACM TechNews, Friday, March 9, 2018



ACM TechNews <technews-editor@acm.org>

Friday, 9 March 2018 at 18:52

To: Nava Tintarev - EWI

[Unsubscribe](#)

[Manage Add-ins...](#)

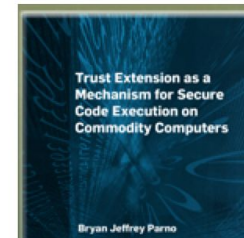
[Click here](#) to view this online



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. Shmoni, 2003.  
“Gender, Genre, and Writing Style in Formal Written Texts,” *Text*, volume 23, number 3, pp. 321–346

## 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.



# Text Classification

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

# Types of machine learning problems



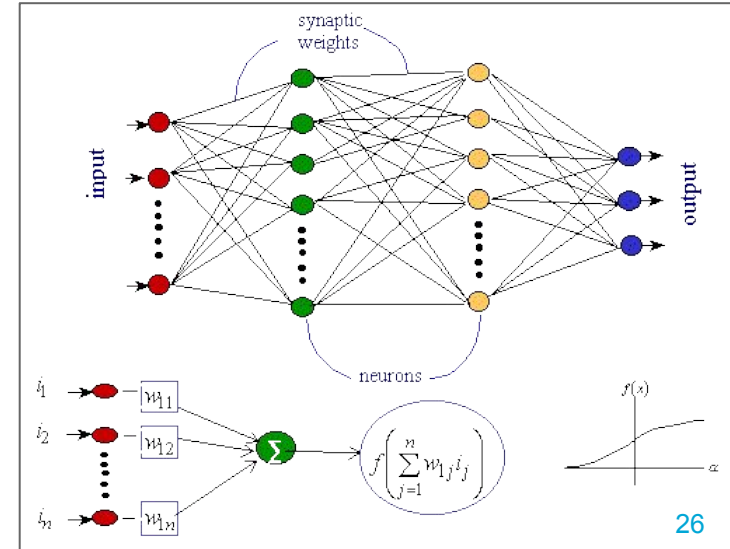
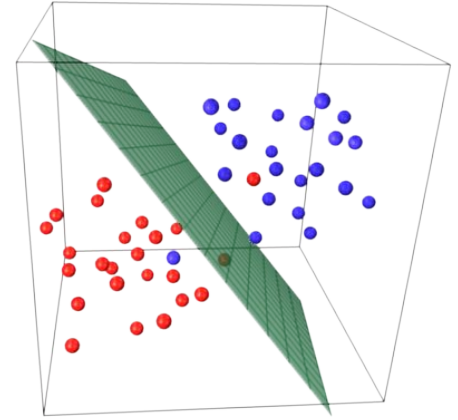
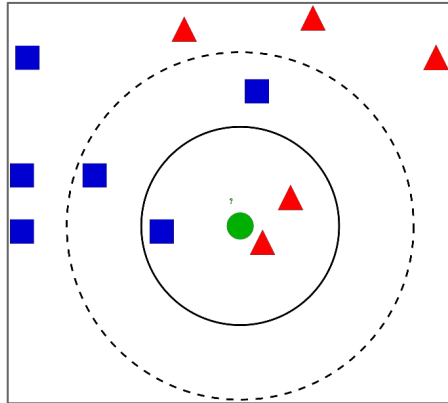
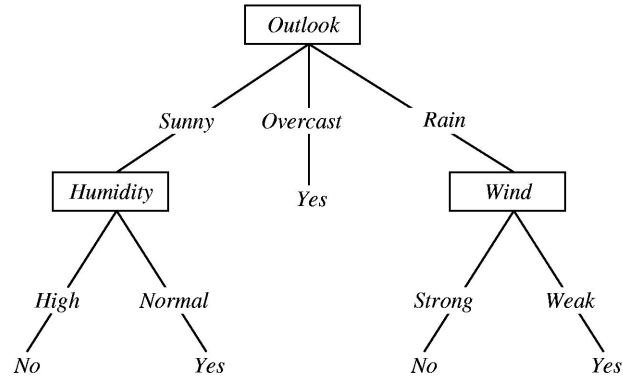


# What is classification?



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

# Classifier examples



# What is classification?

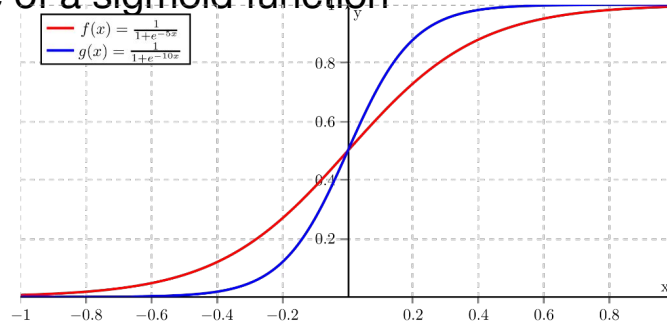
- Different from most **clustering**, where the groups (“clusters”) and their number are not known in advance.
- In **text** classification, data items are textual (e.g., news articles, emails, sentences, queries, etc.) or partly textual (e.g., Web pages)
- The membership of a data item into a class 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



## 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
  - ...

# Fake review detection

How would  
you do it?



Sandra P.  
Wilmington, MA



4/11/2013

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



Laura B.  
Wenham, MA



3/19/2013

Amazing. I highly recommend the El Guapo burrito.



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.



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!



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<sup>2</sup>*

## 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?

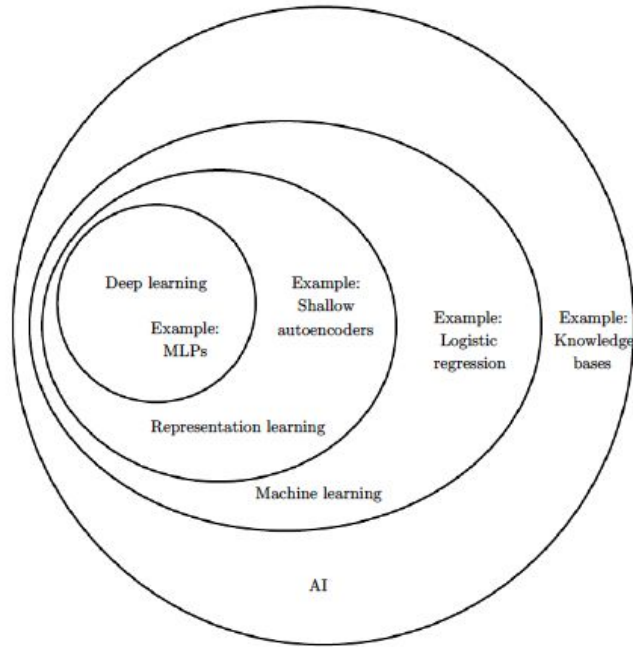


Figure 1.4: A Venn diagram showing how deep learning is a kind of representation learning, which is in turn a kind of machine learning, which is used for many but not all approaches to AI. Each section of the Venn diagram includes an example of an AI technology.

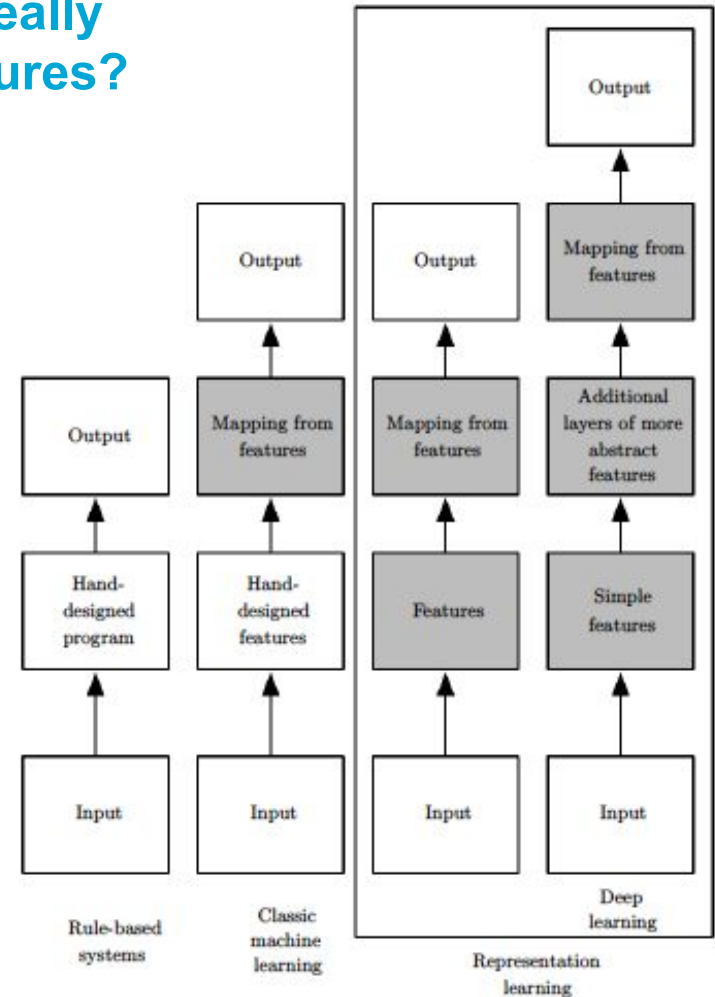


Figure 1.5: Elements showing how the different parts of an AI system relate to each other.

# Fake review detection

How would  
you do it?



0  
10



4/11/2013

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



0  
2



3/19/2013

Amazing. I highly recommend the El Guapo burrito.



0  
2



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



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



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

**How would  
you do it?**

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

# Common Machine Learning techniques



# Machine learning techniques

NB	ME	SVM
----	----	-----

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

77.0	<b>77.7</b>	75.1
80.3	81.0	<b>81.4</b>
81.0	80.1	<b>81.6</b>

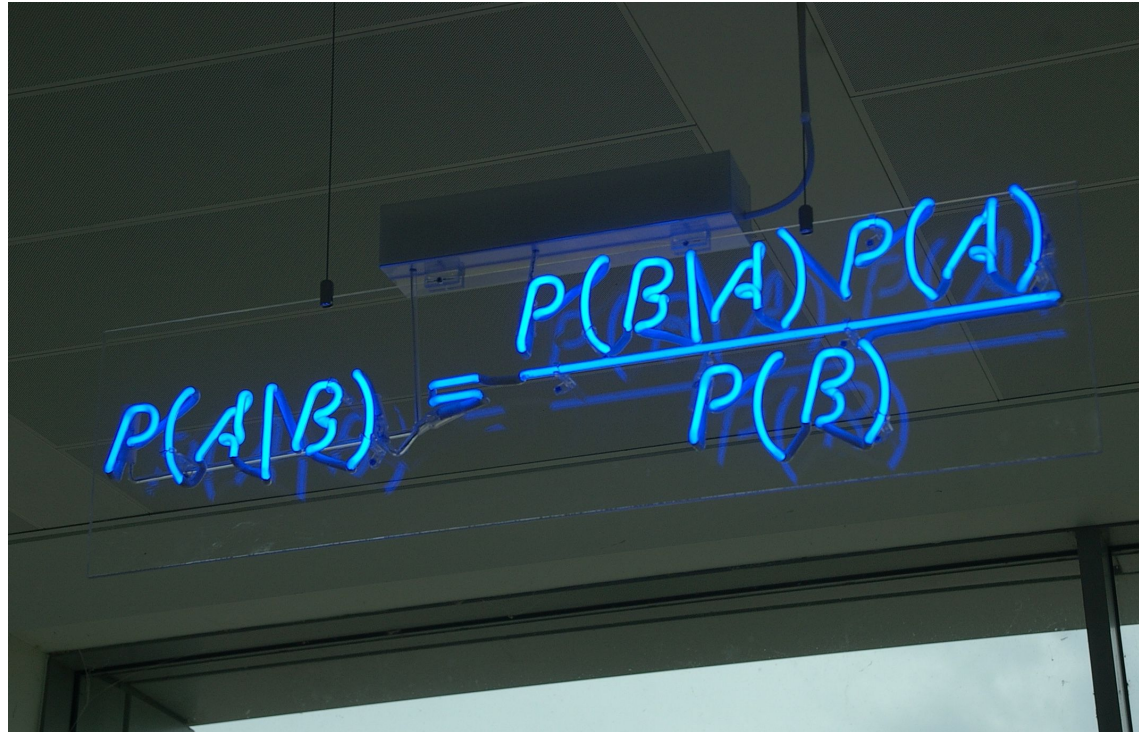
- Naive-Bayes
- Maximal Entropy
- Support Vector Machine

# Common Machine Learning techniques

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



# Naive Bayes



A photograph of a blue neon sign mounted on a ceiling, displaying the Naive Bayes formula. The sign is illuminated and shows the equation  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$  in a stylized, handwritten font. The sign is set against a dark background, and the ceiling structure is visible above it.

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

# Naïve Bayes

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

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

- Training is done by counting and dividing:

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

- 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
  - ...
  - [http://spamassassin.apache.org/tests\\_3\\_3\\_x.html](http://spamassassin.apache.org/tests_3_3_x.html)

## Naive Bayes is Not So Naive

- Very fast learning and testing (basically just count words)
- 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)

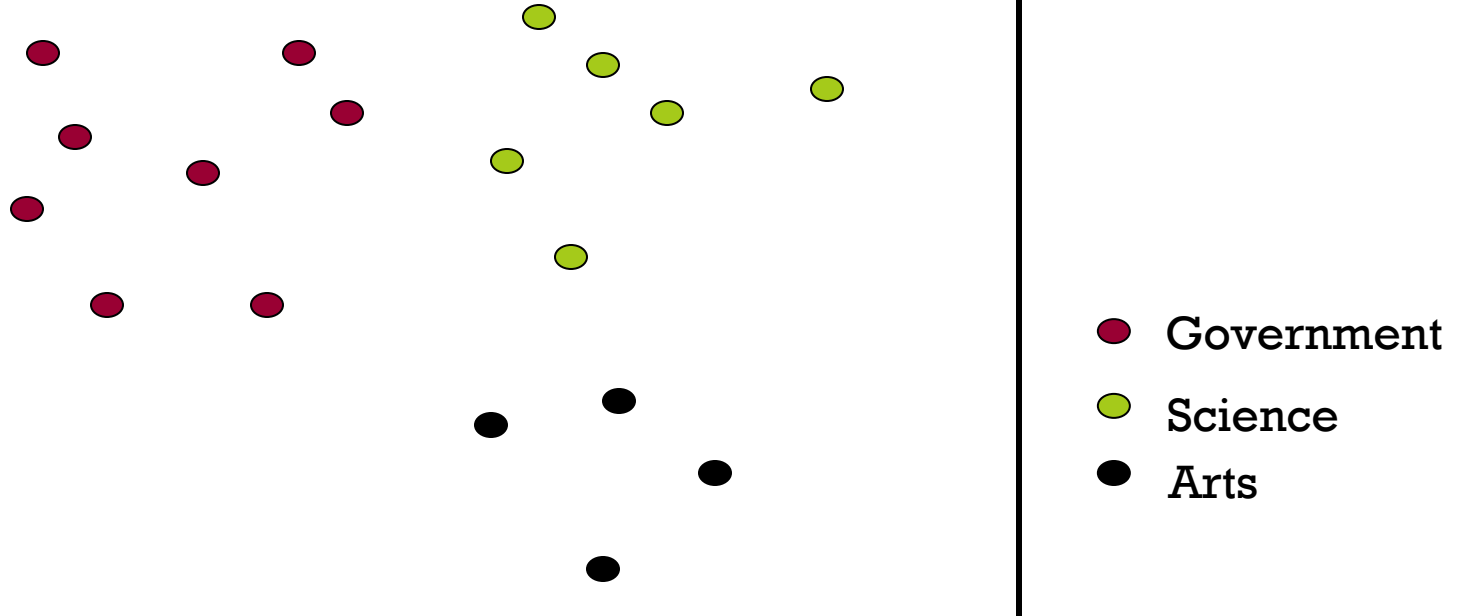
**Premise 1:** Documents in the same class form a contiguous region of space

**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

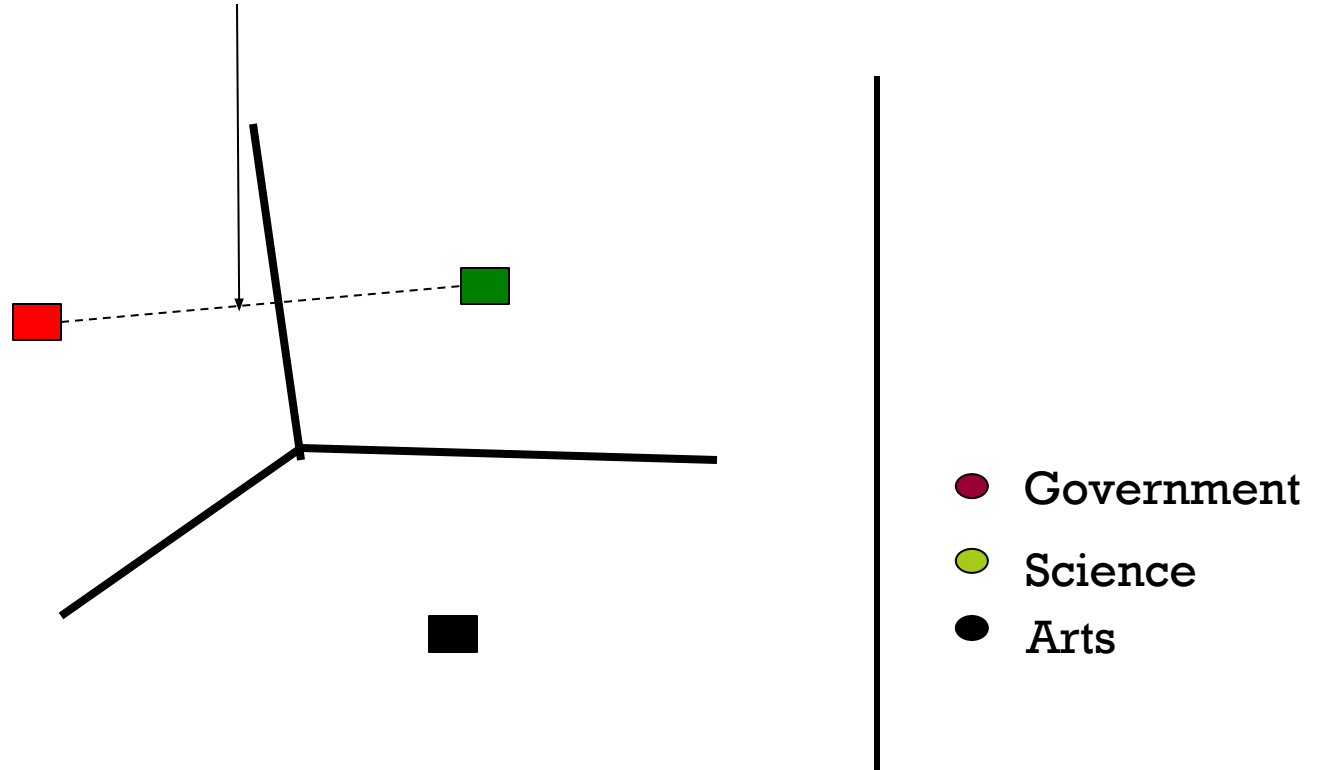


Test Document, of what class?



- Government
- Science
- Arts

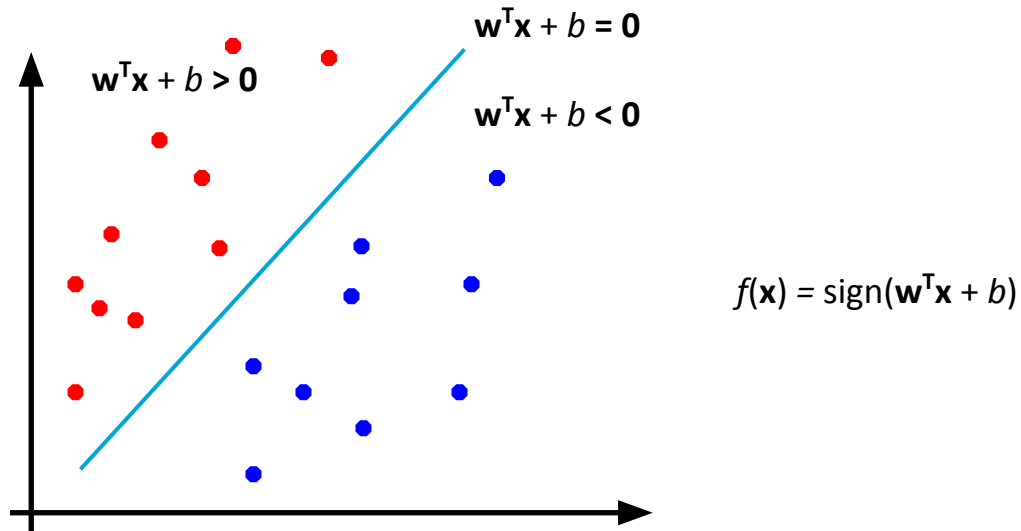
## Test Document = Government



Our focus: how to find good separators

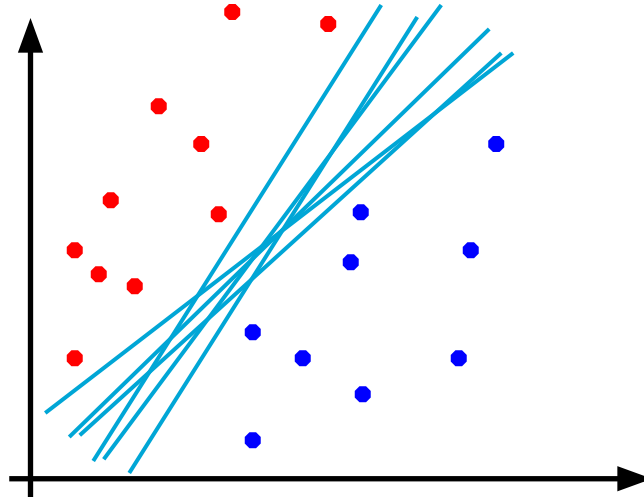
## Linear

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



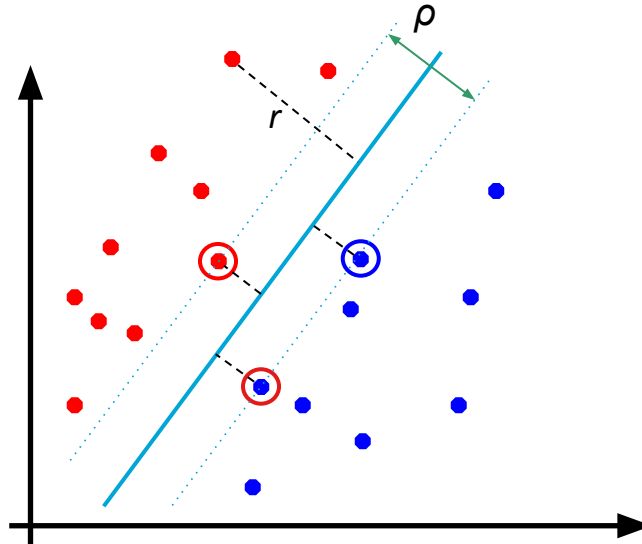
## Linear Separators

- Which of the linear separators is optimal?



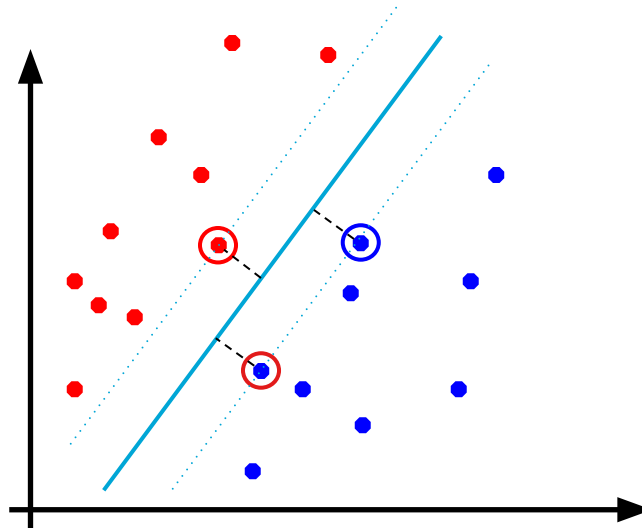
## Classification Margin

- Distance from example  $\mathbf{x}_i$  to the separator is  $r = \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|}$
- Examples closest to the hyperplane are **support vectors**.
- **Margin**  $\rho$  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.



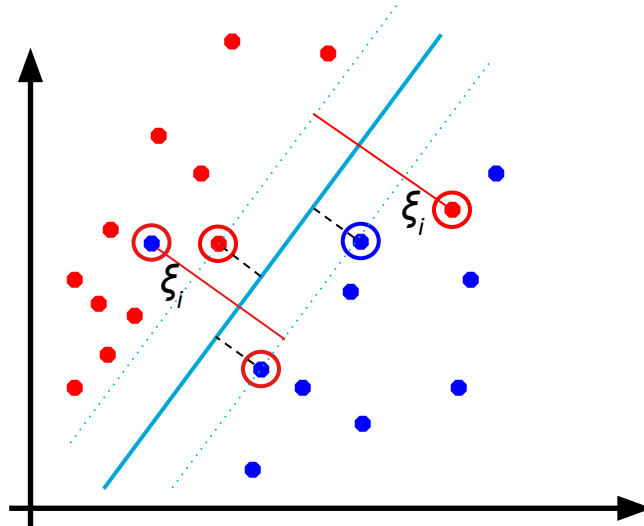
## Optimization Problem

- 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.



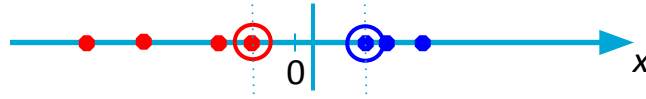
## Soft Margin Classification

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



## Non-linear SVMs

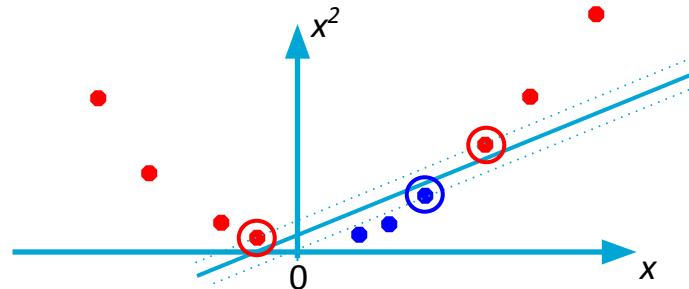
Datasets that are linearly separable with some noise work out great:



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

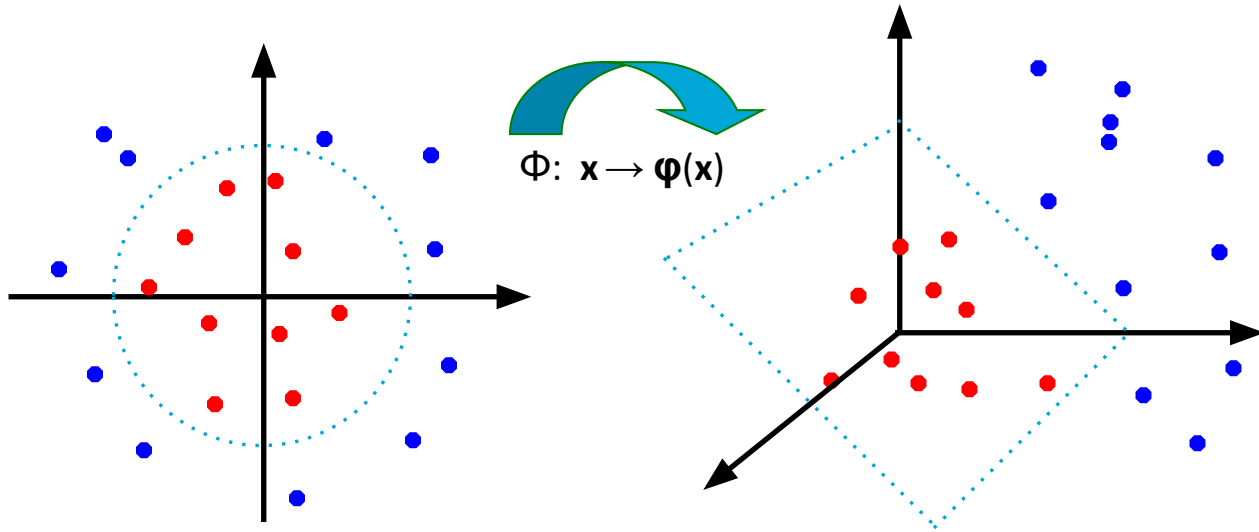


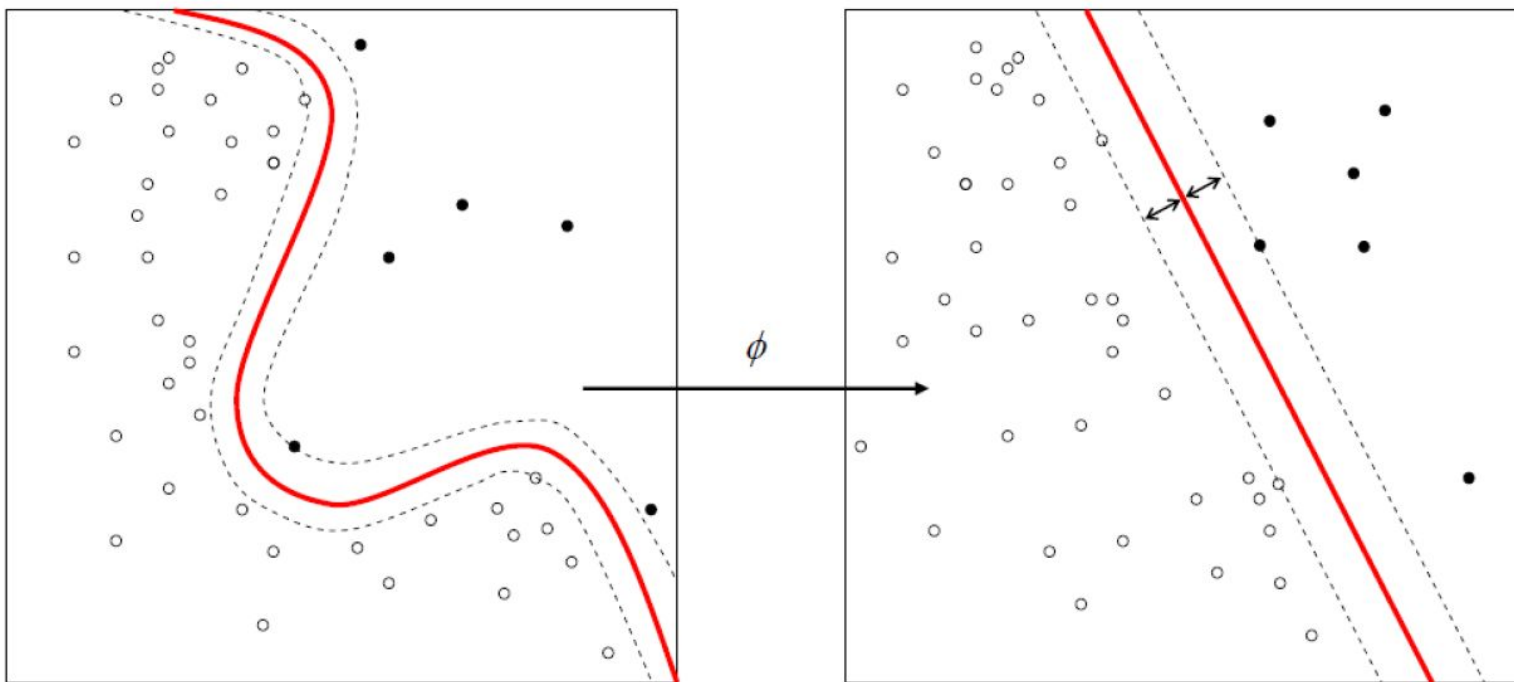
How about... mapping data to a higher-dimensional space:



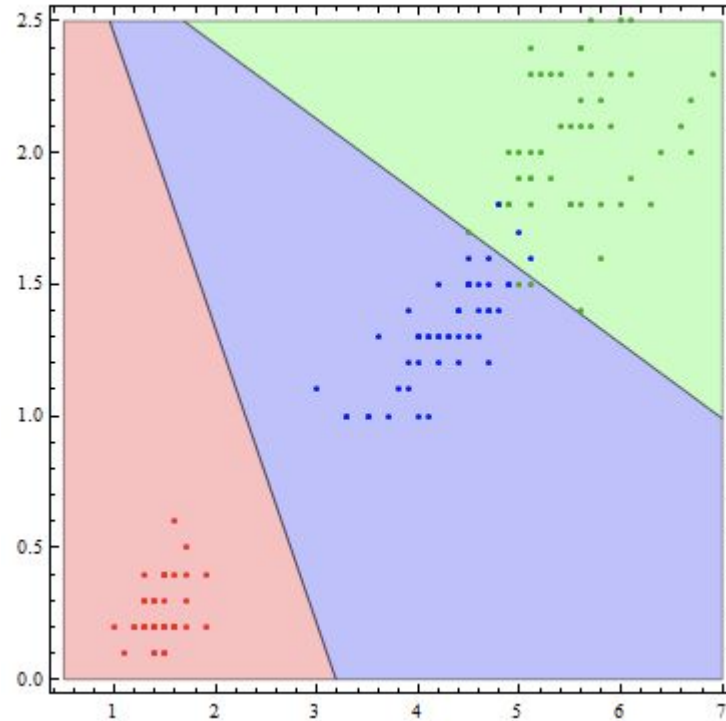
## Non-linear

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





# Maximum Entropy Classifiers



# Maximum Likelihood Estimates

Reminder:

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**

- Intuition: distribution with highest entropy/info fulfilling constraints
- Word usually the same label (low entropy), evenly distributed across labels (high entropy)

# Maximum Likelihood Estimates

- Very similar to Naive Bayes
- Sets the models parameters using search techniques.
  - **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

# 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: a parameter per label (prior), and a 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)

# Exponential Model Likelihood

Maximum (Conditional) Likelihood Models :

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

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

# Maximum Likelihood Estimates

- 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.
    - There are other (good!) ways of discriminating classes – SVMs, boosting, even perceptrons – but these methods are not as trivial to interpret as distributions over classes.
    - Gives a probability distribution over the classes (soft classification).

# The Likelihood Value

- The (log) conditional likelihood according to maxent model is a function of the data and the parameters  $\lambda$ :

$$\log P(C \mid D, \lambda) = \log \prod_{(c,d) \in (C,D)} P(c \mid d, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c \mid d, \lambda)$$

- If there aren't many values of  $c$ , it's easy to calculate:

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

# Common Machine Learning techniques

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

# Discriminative vs. Generative



# More terminology

$$P(c | d)$$

**Discriminative = conditional**

- Estimate the label (c) given instance distribution (d)
- Minimize an upper-bound on training error

$$P(c, d)$$

**Generative = joint**

- Estimate the instance-label distribution

# Discriminative vs. Generative

- A generative *joint* model gives probabilities  $P(d,c)$  and tries to maximize this joint likelihood.
  - It turns out to be trivial to choose weights: just relative frequencies.
- A discriminative *conditional* model gives probabilities  $P(c|d)$ . It takes the data as given and models only the conditional probability of the class.
  - We seek to maximize conditional likelihood.
  - Harder to do
  - More closely related to classification error.



# Generative (joint models)

$$P(c, d)$$

- We have some data  $\{(d, c)\}$  of paired observations  $d$  and hidden classes  $c$ .
- **Joint (generative) models** place probabilities over both observed data and the hidden stuff (generate the observed data from hidden stuff):
  - All the classic Statistical NLP models:
    - $n$ -gram models, Naive Bayes classifiers, hidden Markov models, probabilistic context-free grammars, IBM machine translation alignment models

# Discriminative (conditional) Models

P (c | d)

- Discriminative (conditional) models take the data as given, and put a probability over hidden structure given the data:
  - Logistic regression, conditional loglinear or maximum entropy models, conditional random fields
  - Also, SVMs, (averaged) perceptron, etc. are discriminative classifiers (but not directly probabilistic)

- Linear discriminant function

$$h(\mathbf{x}; \theta_0, \theta_1, \dots, \theta_n) = \theta_0 + \sum_i \theta_i f_i(\mathbf{x})$$

- Logistic form:

$$P(+1 | \mathbf{x}) = \frac{1}{1 + \exp(-h(\mathbf{x}; \theta))}$$

- Multi-class exponential form (maxent):

$$h(\mathbf{x}, y; \theta_0, \theta_1, \dots, \theta_n) = \theta_0 + \sum_i \theta_i f_i(\mathbf{x}, y)$$

$$P(y | \mathbf{x}; \theta) = \frac{\exp h(\mathbf{x}, y; \theta)}{\sum_{y'} \exp h(\mathbf{x}, y'; \theta)}$$

# Discriminative (conditional)

$P(c | d)$

- Give high **accuracy** performance
- They make it easy to incorporate **lots of linguistically important features**
- They allow automatic building of **language independent**, retargetable NLP modules

# Generative (joint)

$$P(c, d)$$

- Easy to train: just count
- Language modeling: probability of observed forms
- More robust
  - Small training sets
  - Label noise
- Full advantage of probabilistic methods

# Practical issues



# Discriminative versus Generative

Conditional classifiers can answer

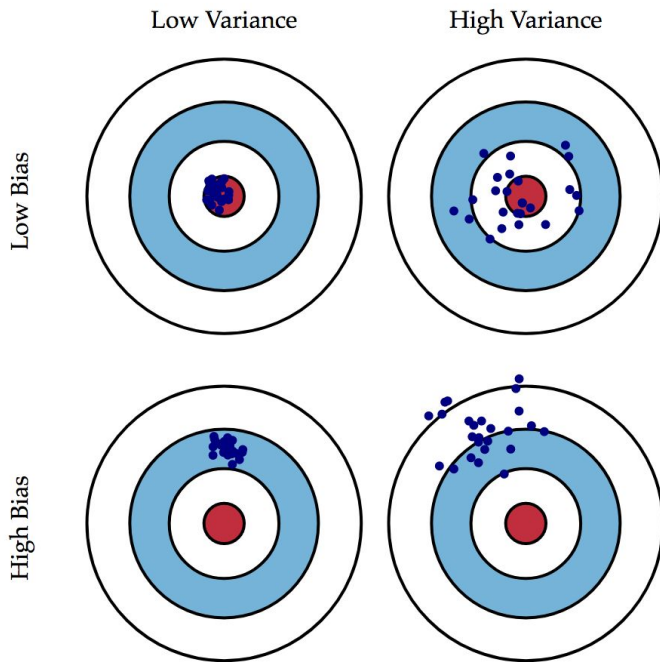
1. What is the most likely label for a given input?
2. How likely is a given label for a given input?

Generative classifiers can also answer

3. What is the most likely input value?
4. How likely is a given input value with a given label?
5. What is the most likely label for an input that might have one of two values (but we don't know which?)

More power,  
more  
parameters,  
same data →  
Less good at  
Q1 and Q2!

# Bias versus variance trade-off



**Error due to Bias:** The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict.

Think over-optimizing precision.

**Error due to Variance:** The error due to variance is taken as the variability of a model prediction for a given data point.

Think over-optimizing recall.

# Very little data?

- Use Naïve Bayes
  - Naïve Bayes is a “high-bias” algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, Expectation maximization over unlabeled documents, ...



# A reasonable amount of data?

- Perfect for all the clever classifiers
  - SVM
  - Logistic Regression
- You can even use user-interpretable decision trees
  - Users like to hack
  - Management likes quick fixes

# A huge amount of data?

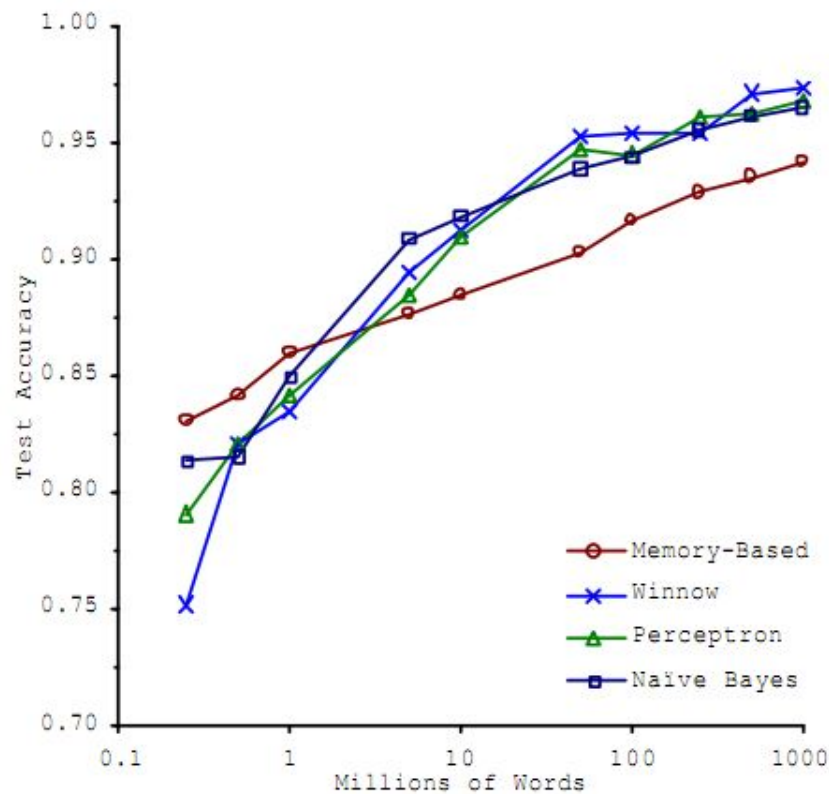
- Can achieve high accuracy!
- At a cost:
  - SVMs (train time) or kNN (test time) can be too slow
  - Regularized logistic regression can be somewhat better
- So Naïve Bayes can come back into its own again!

# How much data?

- **Factor of the number of classes:** There must be  $x$  independent examples for each class, where  $x$  could be tens, hundreds, or thousands (e.g. 5, 50, 500, 5000).
- **Factor of the number of input features:** There must be  $x\%$  more examples than there are input features, where  $x$  could be tens (e.g. 10).
- **Factor of the number of model parameters:** There must be  $x$  independent examples for each parameter in the model, where  $x$  could be tens (e.g. 10).

# Accuracy as a function of data size

With enough data,  
classifier may not  
matter  
...but there is a risk  
over-fitting!



## Real-world systems generally combine:

- Automatic classification
- Manual review of uncertain/difficult/"new" cases

## How to tweak performance

- Domain-specific features and weights: *very* important in real performance
- Sometimes need to collapse terms:
  - Part numbers, chemical formulas, ...
  - But stemming generally doesn't help
- Upweighting: Counting a word as if it occurred twice:
  - title words (Cohen & Singer 1996)
  - first sentence of each paragraph (Murata, 1999)
  - In sentences that contain title words (Ko *et al*, 2002)

# Fake review detection

How would  
you do it?



Sandra P.  
Wilmington, MA



4/11/2013

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



Laura B.  
Wenham, MA



3/19/2013

Amazing. I highly recommend the El Guapo burrito.



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.



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!



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.

# 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)



# Fake Review Detection

## Linguistic features

- Tripadvisor versus AMT: Word distributions are different!
- Yelp versus AMT
  - Small number of words over-used →
  - Lower JS-Div fake reviewers vs non-fake (than AMT)

# Milestones

## Thumbs up?: sentiment classification using machine learning techniques

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

Review P11: handed out March 28, due April 4.

**NLP intermediate project report: due April 5.**

# This week

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

# Credits

Stanford NLP course

Zoltán Szilávik 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](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>

# Questions?

