Machine Learning for Computational Advertising

L1: Basics and Probability Theory

Alexander J. Smola

Yahoo! Labs
Santa Clara, CA 95051
alex@smola.org

UC Santa Cruz, April 2009

Overview

L1: Machine learning and probability theory
Introduction to pattern recognition, classification, regression, novelty detection, probability theory, Bayes rule, density estimation

L2: Instance based learning
Nearest Neighbor, Kernels density estimation, Watson Nadaraya estimator, crossvalidation

L3: Linear models
Hebb's rule, perceptron algorithm, regression, classification, feature maps

L1 Introduction to Machine Learning

Data

- Texts, images, vectors, graphs

What to do with data

- Unsupervised learning (clustering, embedding, etc.)
- Classification, sequence annotation
- Regression
- Novelty detection

Statistics and probability theory

- Probability of an event
- Dependence, independence, conditional probability
- Bayes rule, Hypothesis testing

Density estimation

- empirical frequency, bin counting
- priors and Laplace rule

Outline

1. **Data**
2. Data Analysis
   - Unsupervised Learning
   - Supervised Learning

Data

Vectors

- Collections of features
  e.g. income, age, household size, IP numbers, . . .
- Can map categorical variables into vectors

Matrices

- Images, Movies
- Preferences (see collaborative filtering)

Strings

- Documents (web pages)
- Headers, URLs, call graphs

Structured Objects

- XML documents
- Graphs (instant messaging, link structure, tags, . . .)

Optical Character Recognition

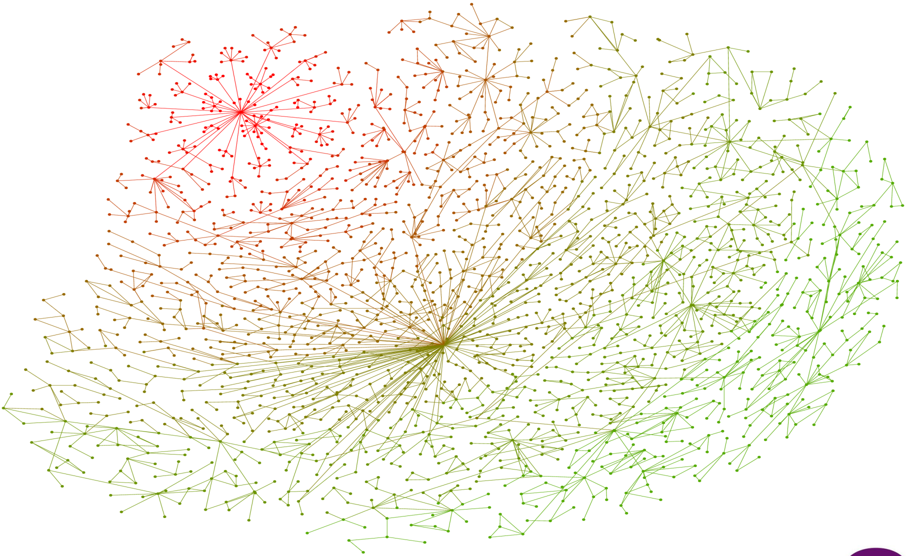
Reuters Database


```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="13522" NEWID="8001">
<DATE>20-MAR-1987 16:54:10.55</DATE>
<TOPICS><D>earn</D></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F
&#22;&#22;&#1;f2479&#31;reute
r f BC-GANTOS-INC-&lt;GTOS>-4TH   03-20 0056</UNKNOWN>
<TEXT>&#2;
<TITLE>GANTOS INC &lt;GTOS> 4TH QTR JAN 31 NET</TITLE>
<DATELINE>    GRAND RAPIDS, MICH., March 20 -
    </DATELINE><BODY>Shr 43 cts vs 37 cts
    Net 2,276,000 vs 1,674,000
    Revs 32.6 mln vs 24.4 mln
    Year
    Shr 90 cts vs 69 cts
    Net 4,508,000 vs 3,096,000
    Revs 101.0 mln vs 76.9 mln
    Avg shrs 5,029,000 vs 4,464,000
    NOTE: 1986 fiscal year ended Feb 1, 1986
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Faces

More Faces

Graphs



Missing Variables

Incomplete Data

- Measurement devices may fail
  E.g. dead pixels on camera, servers fail, forms incomplete, ...
- Measuring things may be expensive
  Need not compute all features for spam if we know
- Data may be censored
  Some users are willing to share more data ...

How to fix it

- Clever algorithms (not this course ...)
- **Simple mean imputation**
  Substitute in the average from other observations
- Works reasonably well — alternatively split features ...

Mini Summary

Data Types

- Vectors (feature sets, bag of words)
- Matrices (photos, dynamical systems, controllers)
- Strings (texts, tags)
- Structured documents (XML, HTML, collections)
- Graphs (web, IM networks)

Problems and Opportunities

- Data may be incomplete (use mean imputation)
- Data may come from different sources (adapt model)
- Data may be biased (e.g. it is much easier to get blood samples from university students for cheap).
- Problem may be ill defined, e.g. “find information.” (get information about what user really needs)
- Environment may react to intervention (butterfly portfolios in stock markets)

Outline

1. Data
2. Data Analysis
   - Unsupervised Learning
   - Supervised Learning

What to do with data

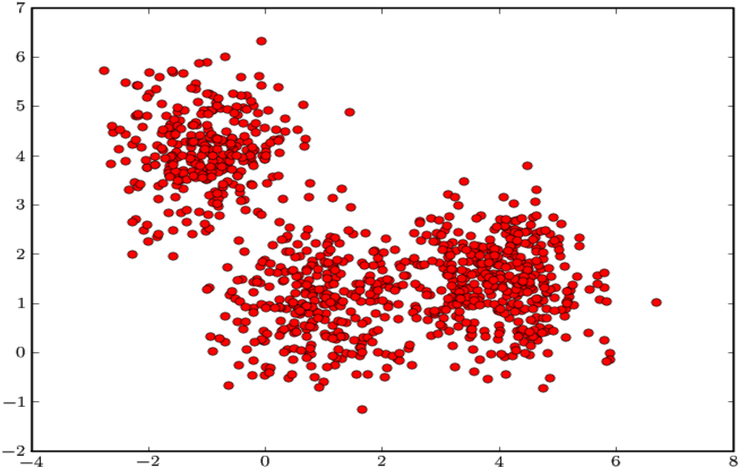
Unsupervised Learning

- Find clusters of the data
- Find low-dimensional representation of the data (e.g. unroll a swiss roll, find structure)
- Find interesting directions in data
- Interesting coordinates and correlations
- Find novel observations / database cleaning

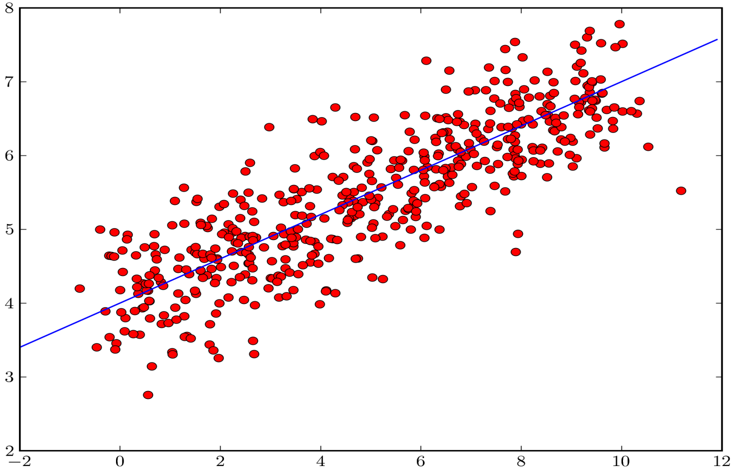
Supervised Learning

- Classification (distinguish apples from oranges)
- Speech recognition
- Regression (advertising price)
- Predict time series (sales forecast)
- Annotate strings (named entities)

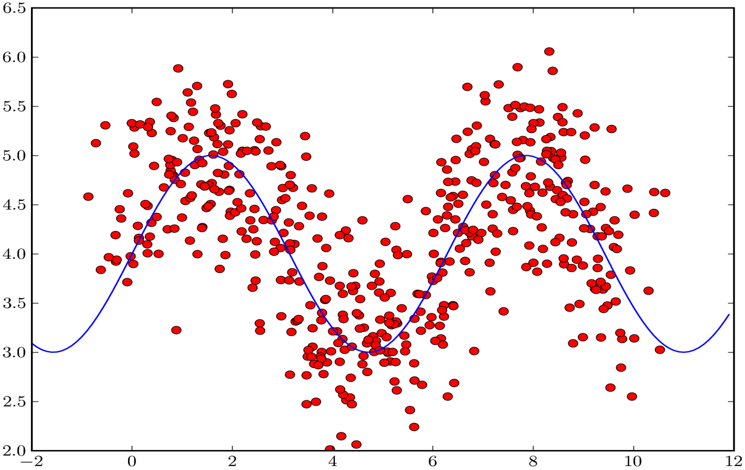
Clustering



Principal Components



Linear Subspace



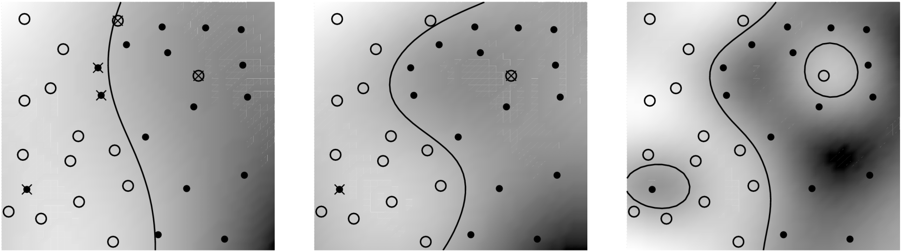
Classification

Data

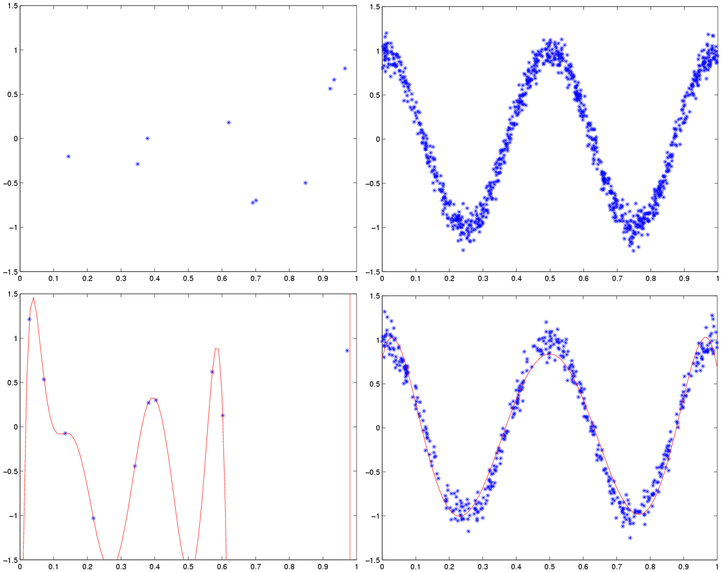
Pairs of observations (x_i, y_i) drawn from distribution
e.g., (query, ad), (credit transactions, fraud), (page, spam)

Goal

Estimate $y \in \{\pm 1\}$ **given** x at a new location. Or find a function $f(x)$ that does the trick.



Regression



Regression

Data

Pairs of observations (x_i, y_i) generated from some joint distribution $\Pr(x, y)$, e.g.,

- (query, ad), bid price
- demographic, expected number of viewers
- web page, quality score (pagerank-ish)

Task

Estimate y, given x, such that some loss $c(x, y, f(x))$ is minimized.

Examples

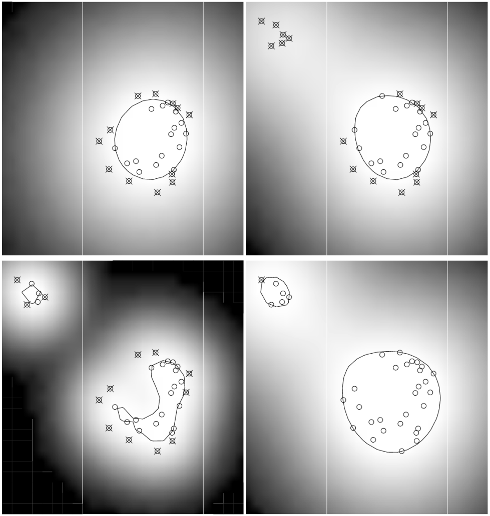
- Quadratic error between y and $f(x)$, i.e. $c(x, y, f(x)) = \frac{1}{2}(y - f(x))^2$.
- Absolute value, i.e., $c(x, y, f(x)) = |y - f(x))|$.

Novelty Detection

Data

Observations x_i from some $\mathrm{P}(x)$, e.g.,

- network usage patterns
- advertisers
- load in server center
- query / news stream



Task

Find unusual events, clean database, typical examples, keep network safe.

Eliza

Rule based AI dialog system, vintage 1970

I am the psychotherapist. Please, describe your problems.

I am here to learn about giving good talks

Why do you say that?

Because I'm interested in learning that

Is it because you are interested in learning that that you came to me?

Yes, indeed.

Earlier you said you are interested in learning that?

Sure, you stupid computer, this is what I want!

Maybe your plans have something to do with this.

Dammit, sure, that's what i'm trying to say.

I don't understand.

Because you're stupid.

Does it bother you that I am stupid?

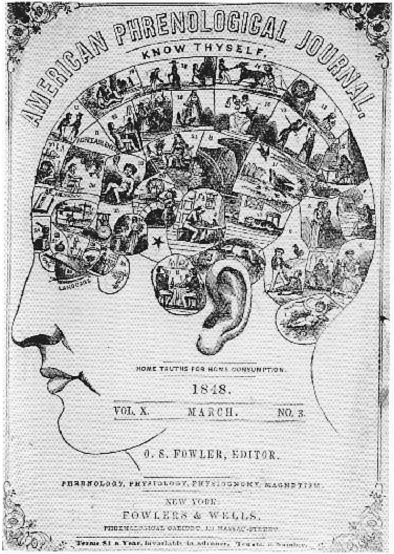
Indeed, it does.

You mentioned that you are interested in learning that?

That's true. And now I want to relax a bit before giving a talk.

Try it out on Emacs using `M-x doctor`.

How the brain doesn’t work



Mini Summary

Structure Extraction

- Clustering
- Low-dimensional subspaces
- Low-dimensional representation of data

Novelty Detection

- Find typical observations (Joe Sixpack)
- Find highly unusual ones (oddball)
- Database cleaning

Supervised Learning

- Regression
- Classification
- Preference relationships (recommender systems)

Statistics and Probability Theory

Why do we need it?

- We deal with **uncertain events**
- Need mathematical formulation for probabilities
- Need to estimate probabilities from data
  (e.g. for coin tosses, we only observe number of heads and tails, not whether the coin is really fair).

How do we use it?

- Statement about probability that an object is an apple (rather than an orange)
- Probability that two things happen at the same time
- Find unusual events (= low density events)
- Conditional events
  (e.g. what happens if A, B, and C are true)

Probability

Basic Idea

We have events in a space of possible outcomes. Then $\Pr(X)$ tells us how likely is that an event $x \in X$ will occur.

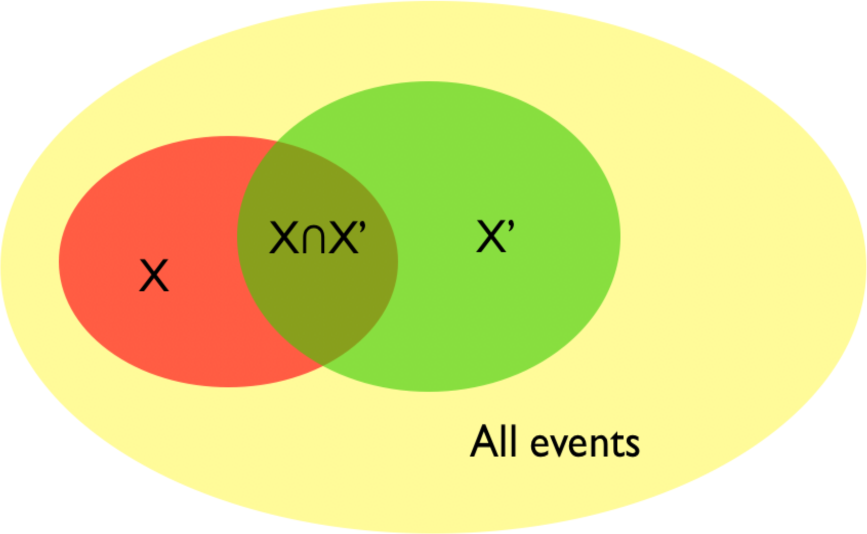
Basic Axioms

- $\Pr(X) \in [0, 1]$ for all $X \subseteq \mathfrak{X}$
- $\Pr(\mathfrak{X}) = 1$
- $\Pr\left(\cup_i X_i\right) = \sum_i \Pr(X_i)$ if $X_i \cap X_j = \emptyset$ for all $i \neq j$

Simple Corollary

$$\Pr(X \cup Y) = \Pr(X) + \Pr(Y) - \Pr(X \cap Y)$$

Example



Multiple Variables

Two Sets

Assume that x and y are drawn from a probability measure on the product space of $\mathcal{X}$ and $\mathcal{Y}$. Consider the space of events $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

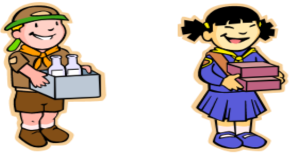
Independence

If x and y are independent, then for all $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$

$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y).$$

Independent Random Variables

Y
Outcome



X
Astrologist's
Prediction

0.25	0.25
0.25	0.25

Dependent Random Variables

Y
Outcome


X
Physician's
Prediction


0.49	0.01
0.01	0.49

Bayes Rule

Dependence and Conditional Probability

Typically, knowing x will tell us something about y (think regression or classification). We have

$$\Pr(Y|X)\Pr(X) = \Pr(Y,X) = \Pr(X|Y)\Pr(Y).$$

- Hence $\Pr(Y,X) \leq \min(\Pr(X),\Pr(Y))$.

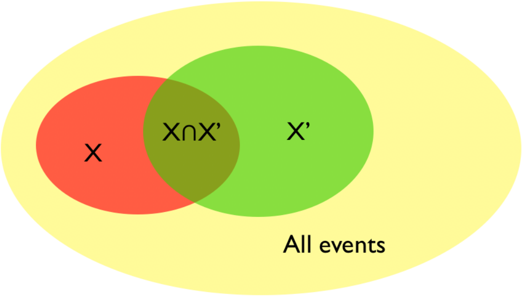
Bayes Rule

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)}.$$

Proof using conditional probabilities

$$\Pr(X,Y) = \Pr(X|Y)\Pr(Y) = \Pr(Y|X)\Pr(X)$$

Example



$$\Pr(X \cap X') = \Pr(X|X')\Pr(X') = \Pr(X'|X)\Pr(X)$$

AIDS Test

How likely is it to have AIDS if the test says so?

- Assume that roughly 0.1% of the population is infected.

$$p(X = \text{AIDS}) = 0.001$$

- The AIDS test reports positive for all infections.

$$p(Y = \text{test positive}|X = \text{AIDS}) = 1$$

- The AIDS test reports positive for 1% healthy people.

$$p(Y = \text{test positive}|X = \text{healthy}) = 0.01$$

We use Bayes rule to infer Pr(AIDS|test positive) via

$$\frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y|X)\Pr(X) + \Pr(Y|\mathcal{X}\backslash X)\Pr(\mathcal{X}\backslash X)}$$

$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Improving Inference — Naive Bayes

Follow up on the AIDS test:

The doctor performs a followup via a conditionally independent test which has the following properties:

- The second test reports positive for 90% infections.
- The AIDS test reports positive for 5% healthy people.

$$\Pr(\mathrm{T1}, \mathrm{T2}|\mathrm{Health}) = \Pr(\mathrm{T1}|\mathrm{Health}) \Pr(\mathrm{T2}|\mathrm{Health}).$$

A bit more algebra reveals (assuming that both tests are independent): $\frac{0.01 \cdot 0.05 \cdot 0.999}{0.01 \cdot 0.05 \cdot 0.999 + 1 \cdot 0.9 \cdot 0.001} = 0.357.$

Conclusion:

Adding extra observations can improve the confidence of the test considerably.

Important Assumption:

We assume that T1 and T2 are **independent** conditioned on health. This is the Naive Bayes classifier.

Different Contexts

Hypothesis Testing:

- Are algorithms *A* or *B* better to solve the problem.
- Can we trust a user (spammer / no spammer)
- Which parameter setting should we use?

Sensor Fusion:

- Evidence from sensors *A* and *B* (AIDS test 1 and 2).
- Different *t*ypes of data (text, images, tags).

More Data:

- We obtain two sets of data — we get more confident
- Each observation can be seen as an additional test

Mini Summary

Probability theory

- Basic tools of the trade
- Use it to model uncertain events

Dependence and Independence

- Independent events don't convey any information about each other.
- Dependence is what we exploit for estimation
- Leads to Bayes rule

Testing

- Prior probability matters
- Combining tests improves outcomes
- Common sense can be misleading

Estimating Probabilities from Data

Rolling a dice:

Roll the dice many times and count how many times each side comes up. Then assign empirical probability estimates according to the frequency of occurrence.

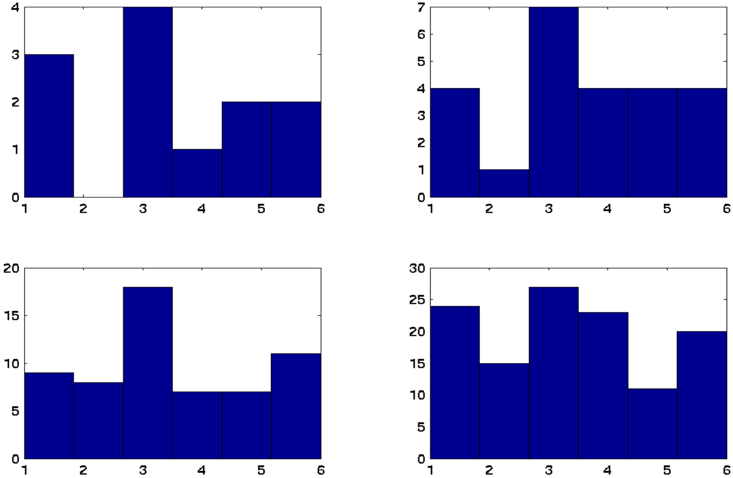
$$\hat{\Pr}(i) = \frac{\#\text{occurrences of } i}{\#\text{trials}}$$

Maximum Likelihood Estimation:

Find parameters such that the observations are *most likely* given the current set of parameters.

This does not check whether the parameters are plausible!

Practical Example



Properties of MLE

Hoeffding's Bound

The probability estimates converge exponentially fast

$$\Pr\{|\pi_i - p_i| > \epsilon\} \leq 2\exp(-2m\epsilon^2)$$

Problem

- For small ϵ this can still take a very long time. In particular, for a fixed confidence level δ we have

$$\delta = 2\exp(-2m\epsilon^2) \Longrightarrow \epsilon = \sqrt{\frac{-\log\delta + \log 2}{2m}}$$

- The above bound holds only for single π_i, **but not uniformly over all i**.

Improved Approach

If we know something about π_i, we should use this extra information: use priors.

Priors to the Rescue

Big Problem

Only sampling *many times* gets the parameters right.

Rule of Thumb

We need at least 10-20 times as many observations.

Conjugate Priors

Often we know what we should expect. Using a conjugate prior helps. We **insert fake additional data** which we assume that it comes from the prior.

Conjugate Prior for Discrete Distributions

- Assume we see u_i additional observations of class i.

$$\pi_i = \frac{\#\text{occurrences of } i + u_i}{\#\text{trials} + \sum_j u_j}.$$

- Assuming that the dice is even, set $u_i = m_0$ for all $1 \leq i \leq 6$. For $u_i = 1$ this is the **Laplace Rule**.

Example: Dice

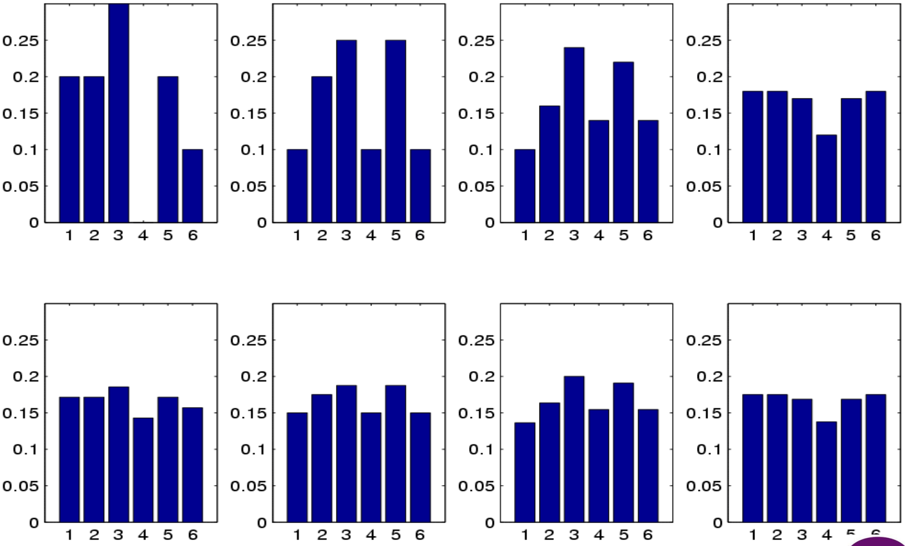
20 tosses of a dice

Outcome	1	2	3	4	5	6
Counts	3	6	2	1	4	4
MLE	0.15	0.30	0.10	0.05	0.20	0.20
MAP ($m_0 = 6$)	0.25	0.27	0.12	0.08	0.19	0.19
MAP ($m_0 = 100$)	0.16	0.19	0.16	0.15	0.17	0.17

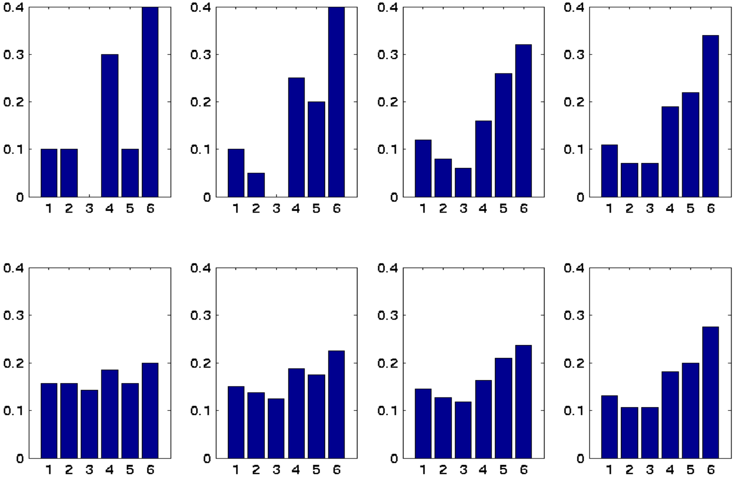
Consequences

- Stronger prior brings the estimate closer to uniform distribution.
- More robust against outliers
- But: Need more data to detect deviations from prior

Honest dice



Tainted dice



Mini Summary

Maximum Likelihood Solution

- Count number of observations per event
- Set probability to empirical frequency of occurrence.

Maximum a Posteriori Solution

- We have a good guess about solution
- Use conjugate prior
- Corresponds to inventing extra data
- Recalibrate probability for additional observations.

Extension

- Works for other estimates, too (means, covariance matrices).

Big Guns: Hoeffding and friends

- Use uniform convergence and tail bounds
- Exponential convergence for fixed scale
- Only sublinear convergence, when fixed confidence.

Summary

Data
Vectors, matrices, strings, graphs, . . .

What to do with data
Unsupervised learning (clustering, embedding, etc.), Classification, sequence annotation, Regression, . . .

Random Variables
Dependence, Bayes rule, hypothesis testing

Estimating Probabilities
Maximum likelihood, convergence, . . .