Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

August 30, 2017

Today:
- Decision trees
- Overfitting
- The Big Picture

Coming soon
- Probabilistic learning
- MLE, MAP estimates

Readings:
Decision trees, overfitting
- Mitchell, Chapter 3

Probabilistic learning
- Estimating Probabilities [Mitchell]
- Andrew Moore's online probability tutorial

Function Approximation:

Problem Setting:
- Set of possible instances X
- Unknown target function $f : X \rightarrow Y$
- Set of function hypotheses $H = \{ h \mid h : X \rightarrow Y \}$

Input:
- Training examples $\{<x^{(i)}, y^{(i)}>\}$ of unknown target function f

Output:
- Hypothesis $h \in H$ that best approximates target function f

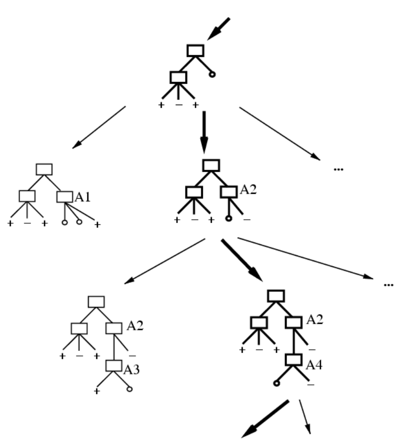
Function Approximation: Decision Tree Learning

Problem Setting:
- Set of possible instances X
  - each instance x in X is a feature vector
    $x = < x_1, x_2 \dots x_n>$
- Unknown target function $f : X \rightarrow Y$
  - Y is discrete valued
- Set of function hypotheses $H=\{ h \mid h : X \rightarrow Y \}$
  - each hypothesis h is a decision tree

Input:
- Training examples $\{<x^{(i)},y^{(i)}>\}$ of unknown target function f

Output:
- Hypothesis $h \in H$ that best approximates target function f

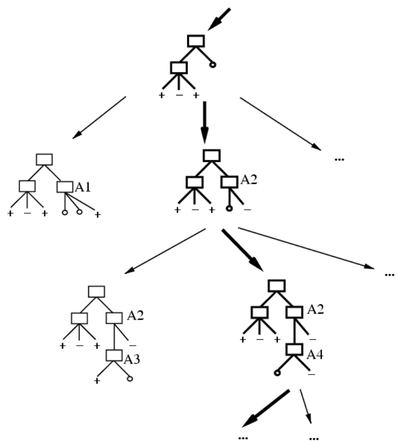
Function approximation as Search for the best hypothesis



- ID3 performs heuristic search through space of decision trees

Function Approximation: The Big Picture

Which Tree Should We Output?



- ID3 performs heuristic search through space of decision trees
- It stops at smallest acceptable tree. Why?

Occam's razor: prefer the simplest hypothesis that fits the data

Why Prefer Short Hypotheses? (Occam's Razor)

Arguments in favor:

Arguments opposed:

Why Prefer Short Hypotheses? (Occam's Razor)

Argument in favor:

- Fewer short hypotheses than long ones

→ a short hypothesis that fits the data is less likely to be a statistical coincidence

Argument opposed:

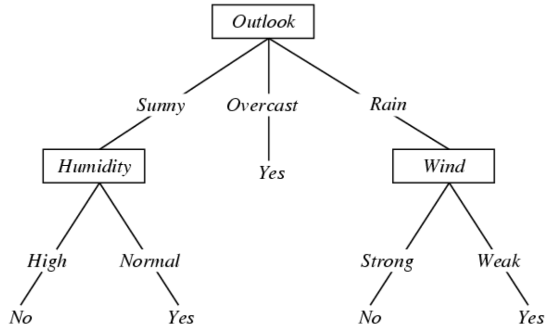
- Also fewer hypotheses containing a prime number of nodes and attributes beginning with "Z"
- What's so special about "short" hypotheses, instead of "prime number of nodes and edges"?

Overfitting in Decision Trees

Consider adding noisy training example #15:

Sunny, Mild, Normal, Strong, PlayTennis=No

What effect on earlier tree?



Overfitting

Consider a hypothesis h and its

- Error rate over training data: $error_{train}(h)$
- True error rate over all data: $error_{true}(h)$

Overfitting

Consider a hypothesis h and its

- Error rate over training data: $error_{train}(h)$
- True error rate over all data: $error_{true}(h)$

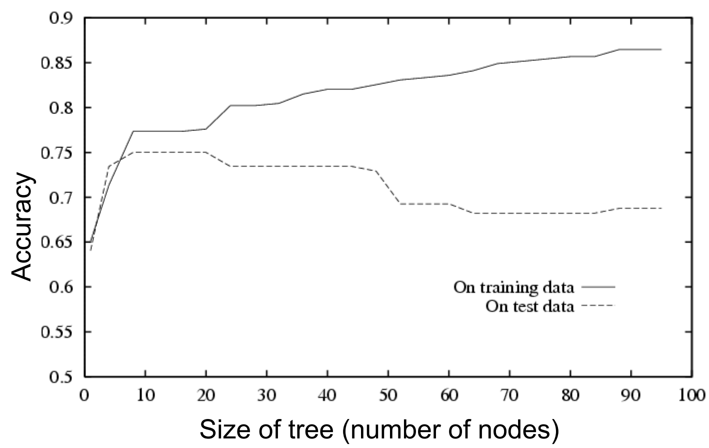
We say h overfits the training data if

$$error_{true}(h) > error_{train}(h)$$

Amount of overfitting =

$$error_{true}(h) - error_{train}(h)$$

Overfitting in Decision Tree Learning



Avoiding Overfitting

How can we avoid overfitting?

- stop growing when data split not statistically significant
- grow full tree, then post-prune

How Can We Avoid Overfitting?

1. stop growing tree when data split is not statistically significant

2. grow full tree, then post-prune

3. learn a collection of trees (decision forest) by randomizing training, then have them vote

Reduced-Error Pruning

Split data into *training* and *validation* set

Learn a tree that classifies *training* set correctly

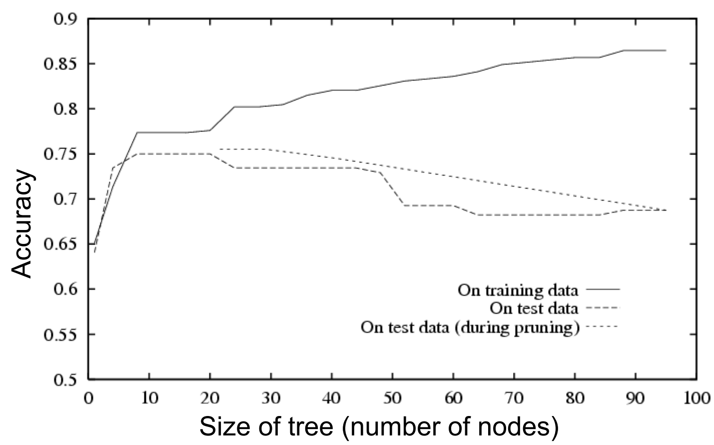
Do until further pruning is harmful:

1. For each non-leaf node, evaluate impact on *validation* set of converting it to a leaf node

2. Greedily select the node that would most improve *validation* set accuracy, and convert it to a leaf

- this produces smallest version of most accurate (over the *validation* set) subtree

Effect of Reduced-Error Pruning



Decision Forests

Key idea:

1. learn a collection of many trees
2. classify by taking a weighted vote of the trees

Empirically successful. Widely used in industry.

- human pose recognition in Microsoft kinect
- medical imaging – cortical parcellation
- classify disease from gene expression data

How to train different trees

1. Train on different random subsets of data
2. Randomize the choice of decision nodes

Decision Forests

Key idea:

1. learn a collection of many trees
2. classify by taking a weighted vote of the trees

more to come

- later lecture on boosting and ensemble methods…

Em

- h
- m
- c

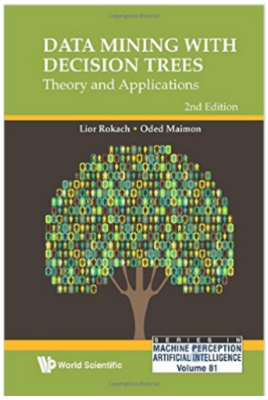
Hov

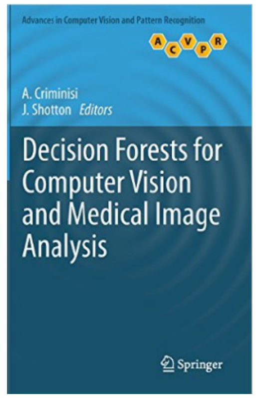
1. Train on different random subsets of data
2. Randomize the choice of decision nodes

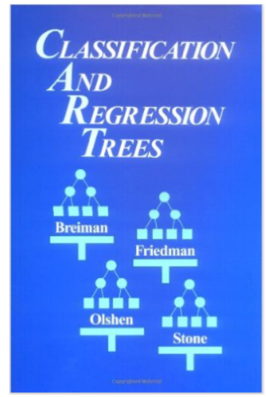
You should know:

- Well posed function approximation problems:
  – Instance space, X
  – Sample of labeled training data { <$x^{(i)}$, $y^{(i)}$>}
  – Hypothesis space, H = { f: X→Y }
- Learning is a search/optimization problem over H
  – Various objective functions to define the goal
    - minimize training error (0-1 loss)
    - minimize validation error (0-1 loss)
    - among hypotheses that minimize error, select smallest (?)
- Decision tree learning
  – Greedy top-down learning of decision trees (ID3, C4.5, ...)
  – Overfitting and post-pruning
  – Extensions… to continuous values, probabilistic classification
  – Widely used commercially: decision *forests*

Further Reading…







Extra slides

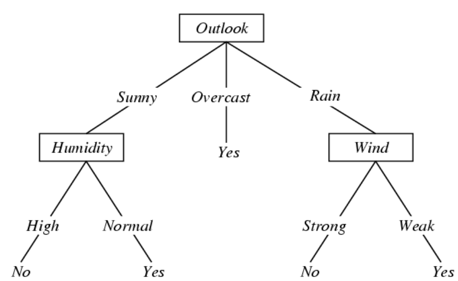
extensions to decision tree learning

Rule Post-Pruning

1. Convert tree to equivalent set of rules
2. Prune each rule independently of others
3. Sort final rules into desired sequence for use

frequently used method (e.g., C4.5)

Converting A Tree to Rules



Unknown Attribute Values

What if some examples missing values of A?
Use training example anyway, sort through tree

- If node n tests A, assign most common value of A among other examples sorted to node n
- assign most common value of A among other examples with same target value
- assign probability p_i to each possible value v_i of A
  - assign fraction p_i of example to each descendant in tree

Classify new examples in same fashion

Questions to think about (1)

- Consider target function f: <x1,x2> → y, where x1 and x2 are real-valued, y is boolean. What is the set of decision surfaces describable with decision trees that use each attribute at most once?

Questions to think about (2)

- ID3 and C4.5 are heuristic algorithms that search through the space of decision trees. Why not just do an exhaustive search?

Questions to think about (3)

- Why use Information Gain to select attributes in decision trees? What other criteria seem reasonable, and what are the tradeoffs in making this choice?

probabilistic function approximation:

instead of $F: X \rightarrow Y$,
learn $P(Y \mid X)$

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment
- Examples
  A = True if a randomly drawn person from our class is female
  A = The hometown of a randomly drawn person from our class
  A = True if two randomly drawn persons from our class have same birthday
- Define P(A) as "the fraction of possible worlds in which A is true" or "the fraction of times A holds, in repeated runs of the random experiment"
  - the set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S

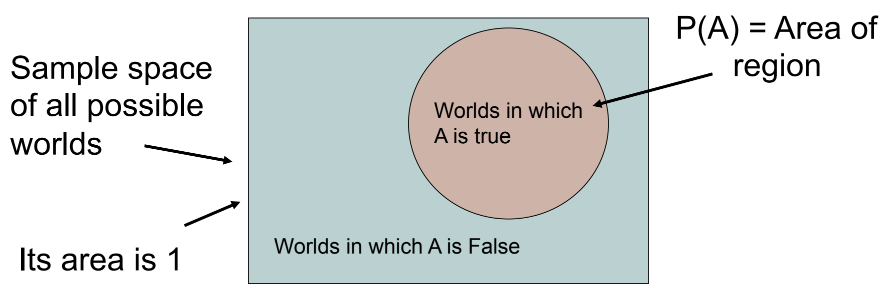
$$A: S \rightarrow \{0,1\}$$

A little formalism

More formally, we have

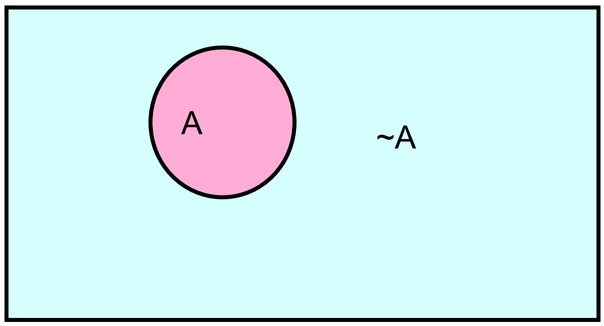
- a <u>sample space</u> S (e.g., set of students in our class)
  - aka the set of possible worlds
- a <u>random variable</u> is a function defined over the sample space
  - Gender: S $\rightarrow$ { m, f }
  - Height: S $\rightarrow$ Reals
- an <u>event</u> is a subset of S
  - e.g., the subset of S for which Gender=f
  - e.g., the subset of S for which (Gender=m) AND (Height > 2m)
- we're often interested in probabilities of specific events
- and of specific events conditioned on other specific events

Visualizing A



Elementary Probability in Pictures

- $P(\sim A) + P(A) = 1$



The Axioms of Probability

- 0 <= P(A) <= 1
- P(True) = 1
- P(False) = 0
- P(A or B) = P(A) + P(B) - P(A and B)

[di Finetti 1931]:

when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms

A useful theorem

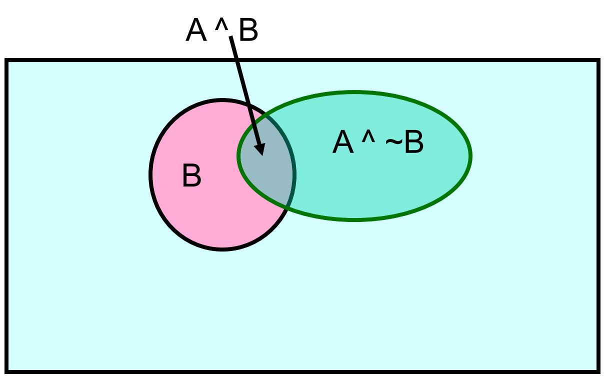
- Axioms: $0 <= P(A) <= 1$, $P(True) = 1$, $P(False) = 0$,
  $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

➔ $P(A) = P(A \wedge B) + P(A \wedge \sim B)$

prove this yourself

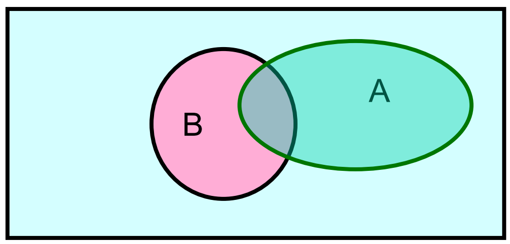
Elementary Probability in Pictures

- $P(A) = P(A \wedge B) + P(A \wedge \sim B)$



Definition of Conditional Probability

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



Definition of Conditional Probability

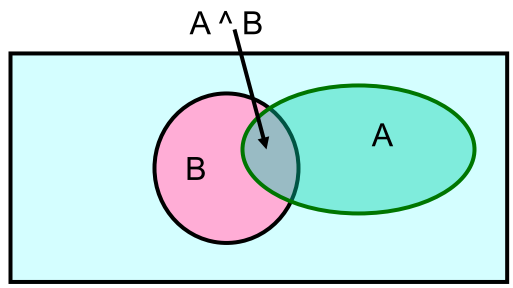
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

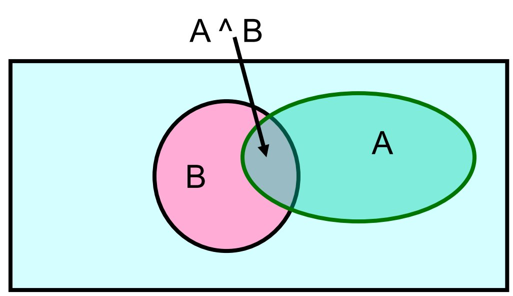
Bayes Rule

- let's write 2 expressions for P(A ^ B)



Bayes Rule

- let's write 2 expressions for P(A ^ B)



$P(A \wedge B) = P(A|B)P(B) = P(B|A)\, P(B)$

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

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Other Forms of Bayes Rule

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

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

Applying Bayes Rule

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

A = you have the flu,   B = you just coughed

Assume:
P(A) = 0.05
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

The Awesome Joint Probability Distribution $P(X_1, X_2, \ldots X_N)$

from which we can calculate
$P(X_1|X_2\ldots X_N)$,
and every other probability we desire
over subsets of $X_1\ldots X_N$

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

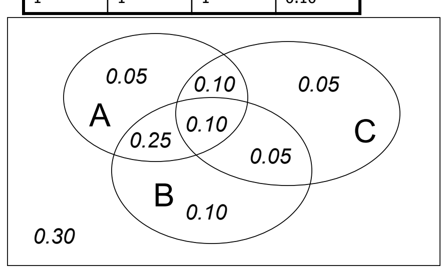
A	**B**	**C**	**Prob**
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	**B**	**C**	**Prob**
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of **any** logical expression involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> → W

Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

sounds like the solution to learning F: X →Y, or P(Y | X).

Are we done?

sounds like the solution to
learning F: X →Y,
or P(Y | X).

Main problem: learning P(Y|X)
can require more data than we have

consider learning Joint Dist. with 100 attributes
# of rows in this table?
# of people on earth?

What to do?

1. Be smart about how we estimate probabilities from sparse data
   – maximum likelihood estimates
   – maximum a posteriori estimates

2. Be smart about how to represent joint distributions
   – Bayes networks, graphical models

1. Be smart about how we estimate probabilities

Estimating Probability of Heads

X=1 X=0

- I show you the above coin X, and ask you to estimate the probability that it will turn up heads (X=1) or tails (X=0)
- You flip it repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_0 times
- Your estimate for $\hat{\theta} = \hat{P}(X = 1)$ is …?

Estimating Probability of Heads

X=1 X=0

- I show you the above coin X, and ask you to estimate the probability that it will turn up heads (X=1) or tails (X=0)
- You flip it repeatedly, observing
  - it turns up heads α_1 times
  - it turns up tails α_0 times

Algorithm 1 (MLE): $$\hat{\theta} = \hat{P}(X = 1) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Estimating θ = P(X=1)

X=1 X=0

Test A:

100 flips: 51 Heads, 49 Tails

Test B:

3 flips: 2 Heads, 1 Tails

Estimating Probability of Heads

When data sparse, might bring in prior assumptions to bias our estimate

- e.g., represent priors by "hallucinating" γ_1 heads, and γ_0 tails, to complement sparse observations

Alg 2 (MAP): $\hat{\theta} = \hat{P}(X=1) = \dfrac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$

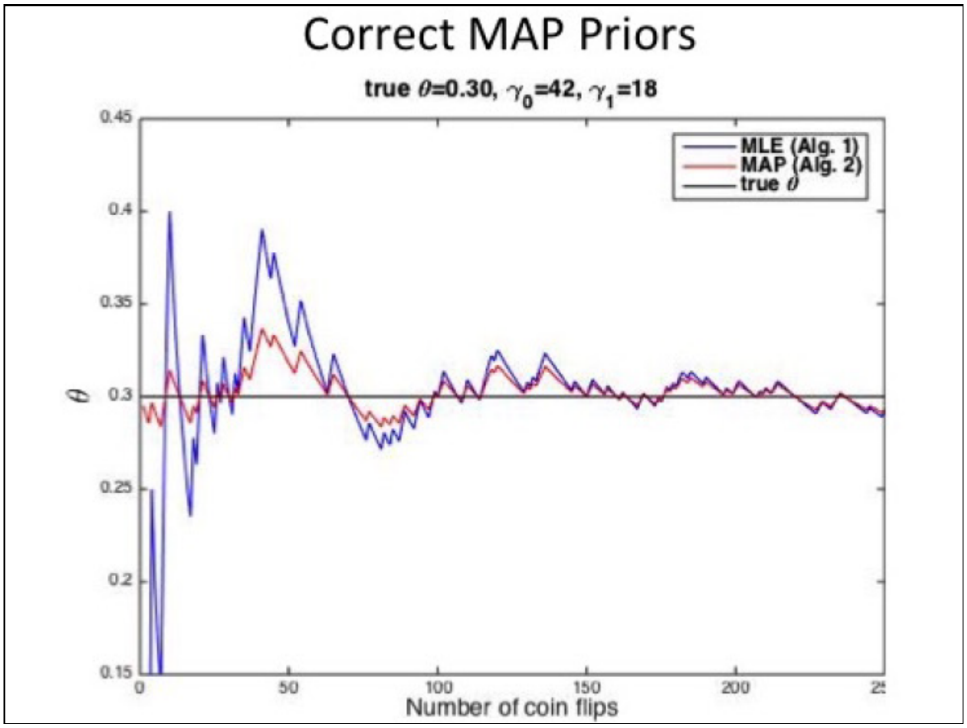
Estimating Probability of Heads

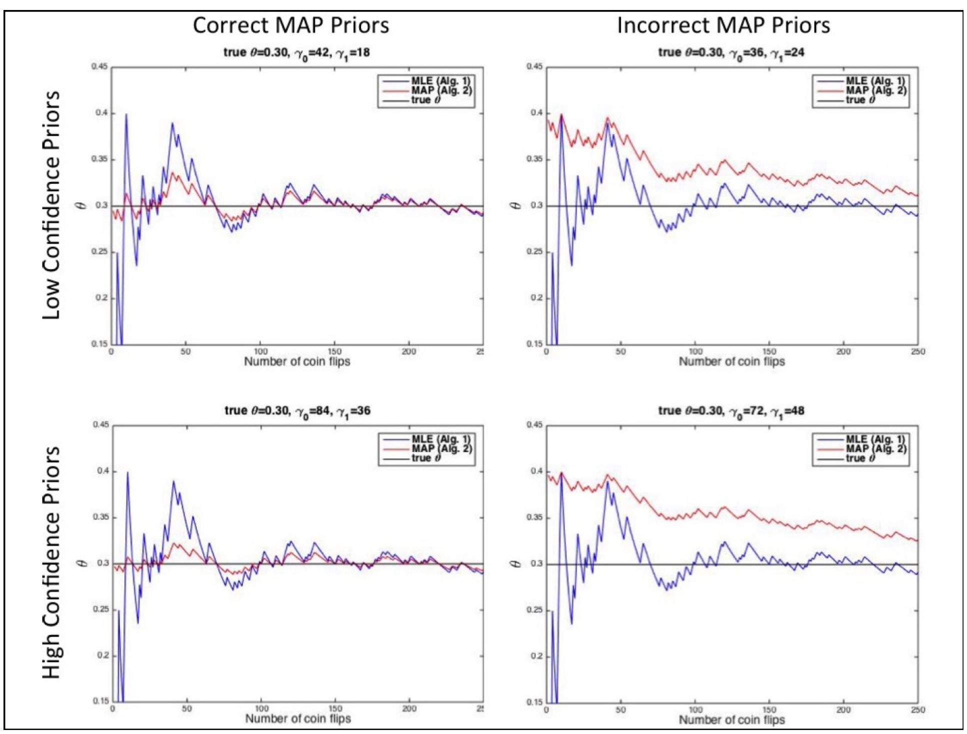
When data sparse, might bring in prior assumptions to bias our estimate

- e.g., represent priors by "hallucinating" γ_1 heads, and γ_0 tails, to complement sparse observations

Alg 2 (MAP): $\hat{\theta} = \hat{P}(X=1) = \dfrac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$

Consider $\gamma_1 = 1 \quad \gamma_0 = 1$
versus $\gamma_1 = 1000 \quad \gamma_0 = 1000$
versus $\gamma_1 = 500 \quad \gamma_0 = 1500$


Correct MAP Priors
true θ=0.30, γ0=42, γ1=18
MLE (Alg. 1)
MAP (Alg. 2)
true θ
θ
Number of coin flips


Correct MAP Priors
Incorrect MAP Priors
Low Confidence Priors
High Confidence Priors
true θ=0.30, γ0=42, γ1=18
true θ=0.30, γ0=36, γ1=24
true θ=0.30, γ0=84, γ1=36
true θ=0.30, γ0=72, γ1=48
MLE (Alg. 1)
MAP (Alg. 2)
true θ
Number of coin flips

Principles for Estimating Probabilities

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and observed data

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta} P(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \\ &= \arg\max_{\theta} P(\mathcal{D} \mid \theta)P(\theta)\end{aligned}$$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize **P(data | θ)**
- result in our case: $\hat{\theta}^{MLE} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$

Principle 2 (maximum a posteriori probability):

- choose parameters θ that maximize **P(θ | data)**
- result in our case:

$$\hat{\theta}^{MAP} = \frac{\alpha_1 + \#\text{hallucinated\_1s}}{(\alpha_1 + \#\text{hallucinated\_1s}) + (\alpha_0 + \#\text{hallucinated\_0s})}$$

Maximum Likelihood Estimation

given data D, choose θ that maximizes P(D | θ)

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

Data D:

$P(D|\theta) =$

Maximum Likelihood Estimation

given data D, choose θ that maximizes P(D | θ)

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

Data D: < 1 0 0 1 1 >

$$P(D|\theta) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$
$$= \theta^{\alpha_1} \cdot (1-\theta)^{\alpha_0}$$

Flips are independent, identically distributed 1's and 0's, producing α_1 1's, and α_0 0's

Now solve for:

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta)$$
$$= \arg\max_{\theta} P(\alpha_1, \alpha_0|\theta)$$
$$= \arg\max_{\theta} \ \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta} = \arg\max_{\theta} \; \ln P(D|\theta)$$

$$= \arg\max_{\theta} \; \ln \left[\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right]$$

- Set derivative to zero: $\boxed{\frac{d}{d\theta} \ln P(\mathcal{D} \mid \theta) = 0}$

hint: $\frac{\partial \ln \theta}{\partial \theta} = \frac{1}{\theta}$

Summary: Maximum Likelihood Estimate for Bernoulli random variable

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros (Binomial)

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \text{argmax}_\theta \, P(D|\theta) = \frac{\alpha_1}{\alpha_1+\alpha_0}$$

Principles for Estimating Probabilities

Principle 1 (maximum likelihood):

- choose parameters θ that maximize P(data | θ)

Principle 2 (maximum a posteriori prob.):

- choose parameters θ that maximize

$$P(\theta \mid \text{data}) = \frac{P(\text{data} \mid \theta)\, P(\theta)}{P(\text{data})}$$

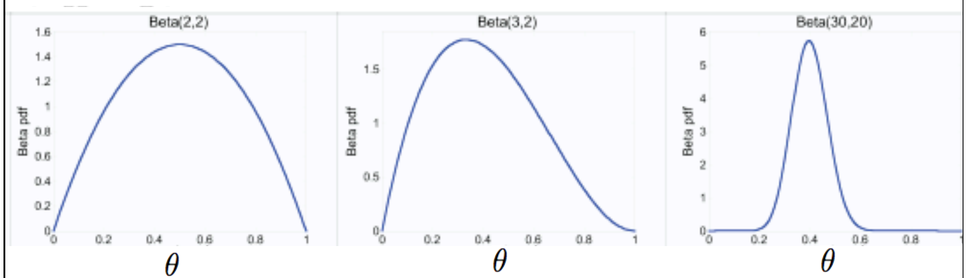
Beta prior distribution : P(θ)

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

Beta prior distribution – P(θ)

- $$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$



Summary: Maximum a Posteriori (MAP) Estimate for Bernoulli random variable

X=1 X=0

P(X=1) = θ
P(X=0) = 1-θ
(Bernoulli)

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \propto P(D|\theta)P(\theta) \sim Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

and MAP estimate is therefore

$$\hat{\theta}^{MAP} = \frac{\alpha_H + \beta_H - 1}{(\alpha_H + \beta_H - 1) + (\alpha_T + \beta_T - 1)}$$

Maximum a Posteriori (MAP) Estimate for random variable with k possible outcomes

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1 - 1}\ \theta_2^{\beta_2 - 1} \ldots \theta_k^{\beta_k - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \mathrm{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \propto P(D|\theta)P(\theta) \sim \mathrm{Dirichlet}(\alpha_1 + \beta_1, \ldots, \alpha_k + \beta_k)$$

and MAP estimate is therefore

$$\hat{\theta}_i^{MAP} = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k} (\alpha_j + \beta_j - 1)}$$

Some terminology

- Likelihood function: P(data | θ)
- Prior: P(θ)
- Posterior: P(θ | data)

- Conjugate prior: P(θ) is the conjugate prior for likelihood function P(data | θ) if the forms of P(θ) and P(θ | data) are the same.
  - Beta is conjugate prior for Bernoulli, Binomial
  - Dirichlet is conjugate prior for Multinomial

You should know

- Probability basics
  - random variables, conditional probs, …
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution
- Estimating parameters from data
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – Bernoulli, Binomial, Beta, Dirichlet, …
  - conjugate priors

Extra slides

Independent Events

- Definition: two events A and B are *independent* if $P(A \wedge B)=P(A)*P(B)$
- Intuition: knowing A tells us nothing about the value of B (and vice versa)

Picture "A independent of B"

Expected values

Given a discrete random variable X, the expected value of X, written E[X] is

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

Example:

X	P(X)
0	0.3
1	0.2
2	0.5

Expected values

Given discrete random variable X, the expected value of X, written E[X] is

$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

We also can talk about the expected value of functions of X

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x) P(X = x)$$

Covariance

Given two discrete r.v.'s X and Y, we define the covariance of X and Y as

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., X=GENDER, Y=PLAYS_FOOTBALL

or X=GENDER, Y=LEFT_HANDED

Remember: $$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$

Conjugate priors

- $P(\theta)$ and $P(\theta|D)$ have the same form

Eg. 1 Coin flip problem

Likelihood is ~ Binomial

$$P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

If prior is Beta distribution,

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

Then posterior is Beta distribution

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

[A. Singh]

Conjugate priors

- P(θ) and P(θ|D) have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2}\ldots\theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k}\theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

31

[A. Singh]

Dirichlet distribution

- number of heads in N flips of a two-sided coin
  - follows a *binomial distribution*
  - Beta is a good prior (conjugate prior for binomial)
- what it's not two-sided, but k-sided?
  - follows a *multinomial distribution*
  - *Dirichlet* distribution is its conjugate prior

$$P(\theta_1, \theta_2, \ldots\theta_K) = \frac{1}{B(\alpha)}\prod_i^K \theta_i^{(\alpha_1 - 1)}$$

Lejeune Dirichlet

Johann Peter Gustav Lejeune Dirichlet

Born	13 February 1805 Düren, French Empire
Died	5 May 1859 (aged 54) Göttingen, Hanover
Residence	Germany
Nationality	German
Fields	Mathematician
Institutions	University of Berlin University of Breslau University of Göttingen
Alma mater	University of Bonn
Doctoral advisor	Simeon Poisson Joseph Fourier
Doctoral students	Ferdinand Eisenstein Leopold Kronecker Rudolf Lipschitz Carl Wilhelm Borchardt
Known for	Dirichlet function Dirichlet eta function