Tutorial

Concepts in Machine Learning and Bayesian Inference

Gustavo Lacerda

What is Machine Learning

- Learning patterns in the world
- Drawing conclusions from data ("problem of induction")
- **Goal:** build a prediction machine

- The scope is the same as Statistics (some say that ML = statistical computing), and Data Mining

- The cultures are different:
  - ML came from AI and Engineering (~1980s)
  - Statistics came from math and applications (1900s?)
  - Data Mining came from Databases, and tends to focus more on huge data sets, and HCI

- Regardless, ML would not exist without computers.

Machine Learning (some types)

Supervised Learning: learn to make predictions based on an input (e.g. classification, regression)

Unsupervised Learning: learn to make predictions *without* any input. (e.g. density modeling, clustering)

Inductive Logic Programming: learn logical rules about the world.

Online Learning: learn as you go.

Active Learning: collect the data that would be most useful to your learning goal.

Reinforcement Learning: learning to behave so as to maximize rewards over the long run.

This talk is about learning to *predict*. If you want to learn to be *happy*, see Hamid's RL tutorial.

Statistical Inference (classical hypothesis test)

Let's play with a coin.

Suppose you toss a coin 20 times... you'd expect it to come up heads ~10 times...

but you only see heads 5 times!

Should you be suspicious of this coin?

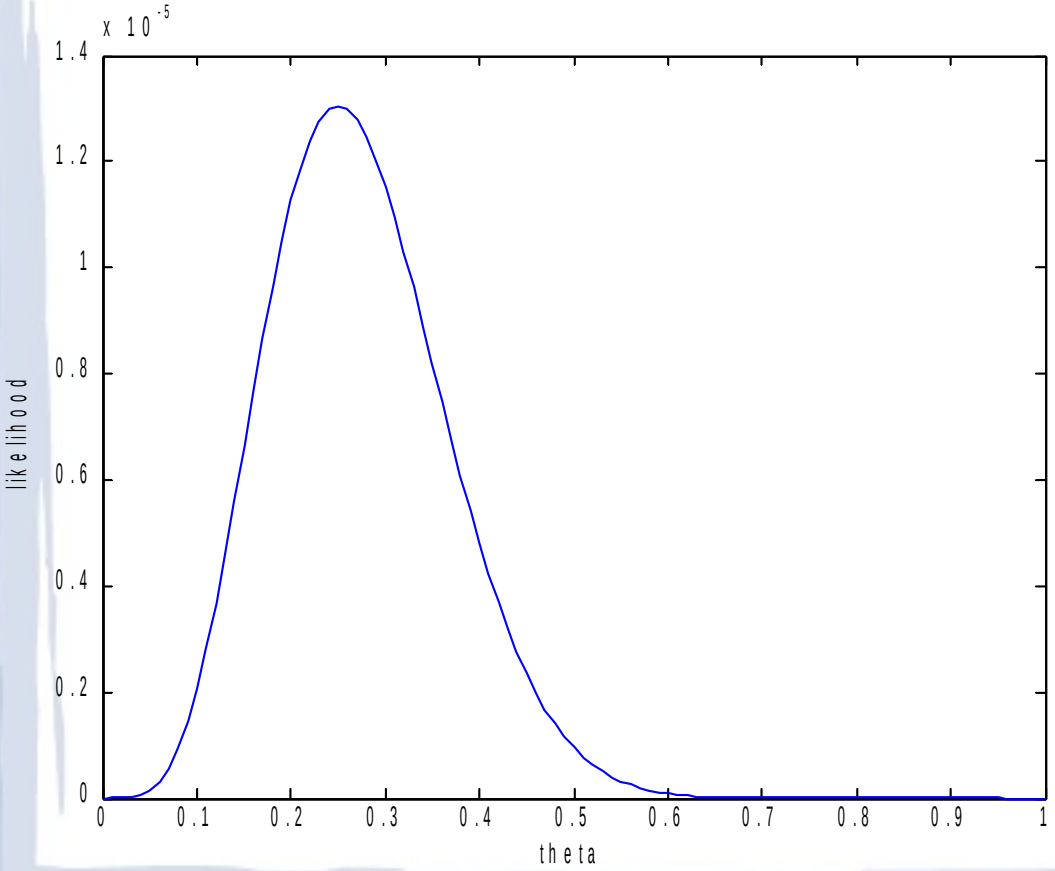
A classical statistician will usually start from a **null hypothesis** (say $\theta = 0.5$), and compute the probability of getting data as extreme as this (known as the **p-value**).

If the p-value is small enough, reject.

Statistical Inference (max-likelihood)

Let's define the **likelihood function**: $f(\theta) = P(D|\theta)$

assuming IID, $f(\theta) = \theta^5 (1-\theta)^{15}$



θ a.k.a. "the probability of heads", is called the **parameter** of this model.

$$\theta_{MLE} = 0.25$$

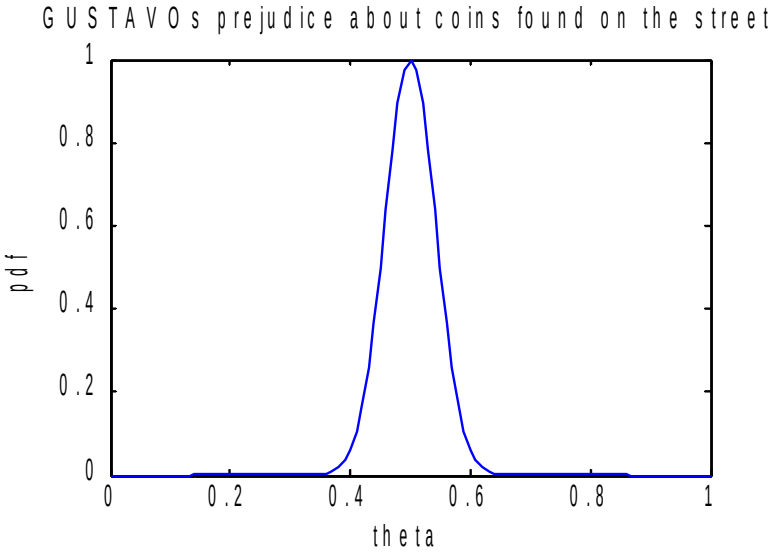
This is called **estimating**, or **fitting** the parameter to the data.

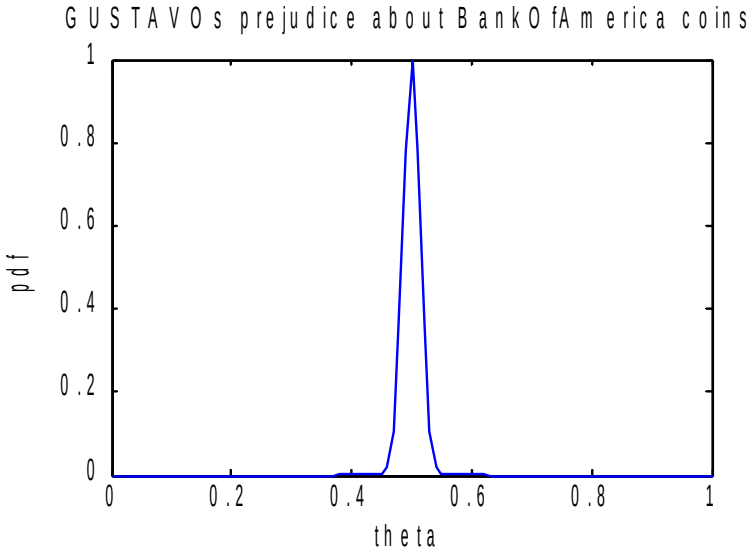
Statistical Inference (Bayesian)

- Suppose you found one coin on the street, and another coin was given to you by Bank of America: shouldn't this matter?

- Everyone agrees that the data is random. “Bayesians” are those who assign a probability distribution to the parameters too.

- The **prior distribution** of parameters represents what you believe about them *before* seeing the data. (a.k.a. your prejudice)





Statistical Inference (Bayesian)

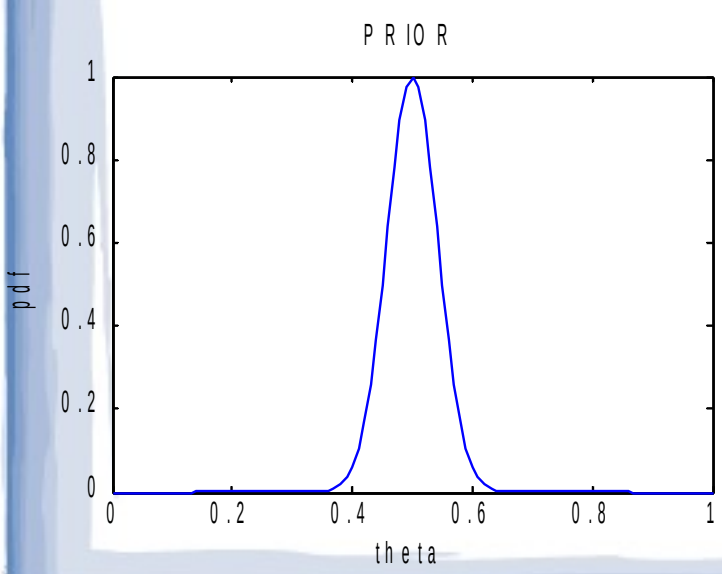
- Let's take a coin that I found on the street.
- Bayes' rule tells us how to **update** our prior with the evidence:

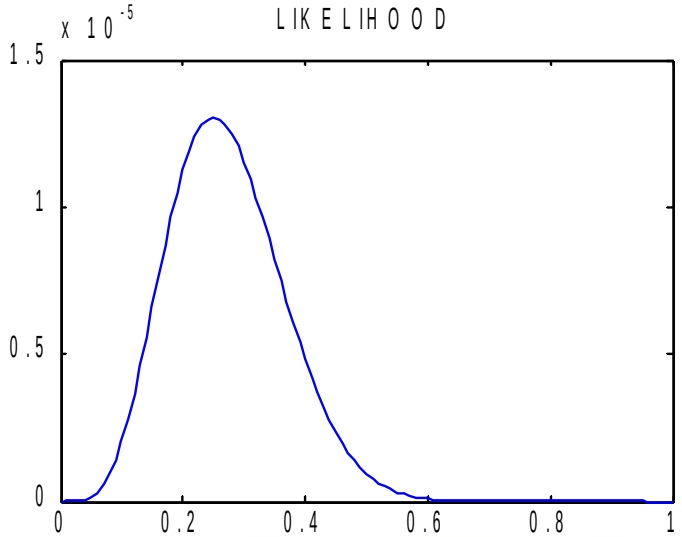
$$P(\theta|D) = P(\theta) * P(D|\theta) / P(D)$$

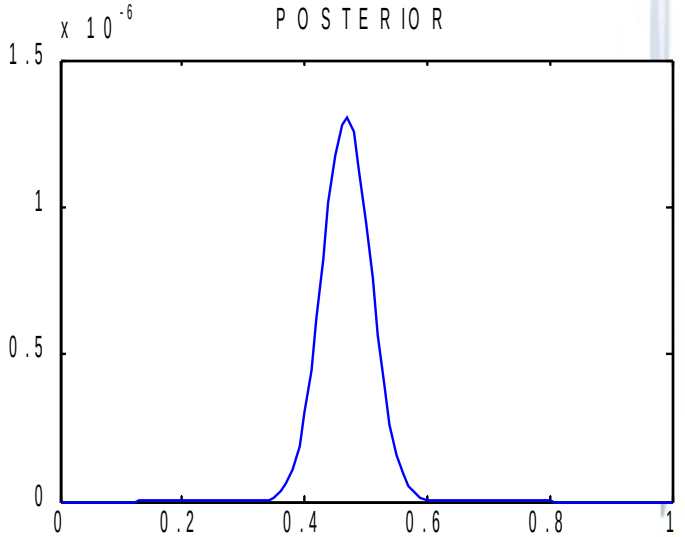
posterior = prior * likelihood (normalized)

- $\theta_{MAP} \sim= .45$

(warning: posteriors not normalized)







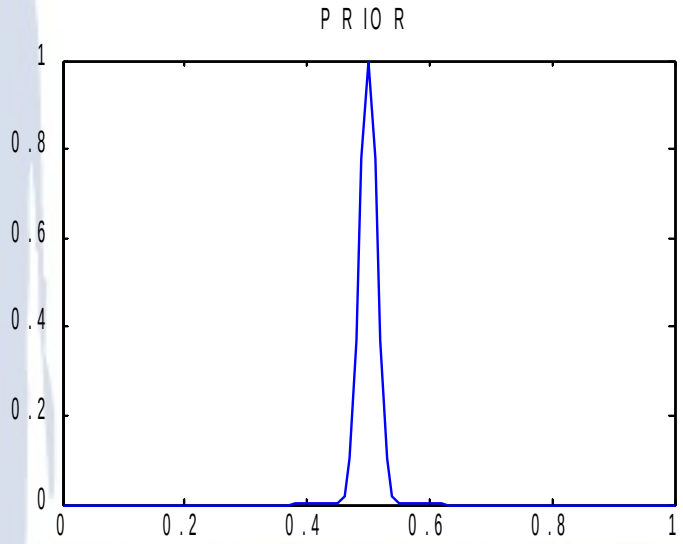
Statistical Inference (Bayesian, with a narrow-minded prior)

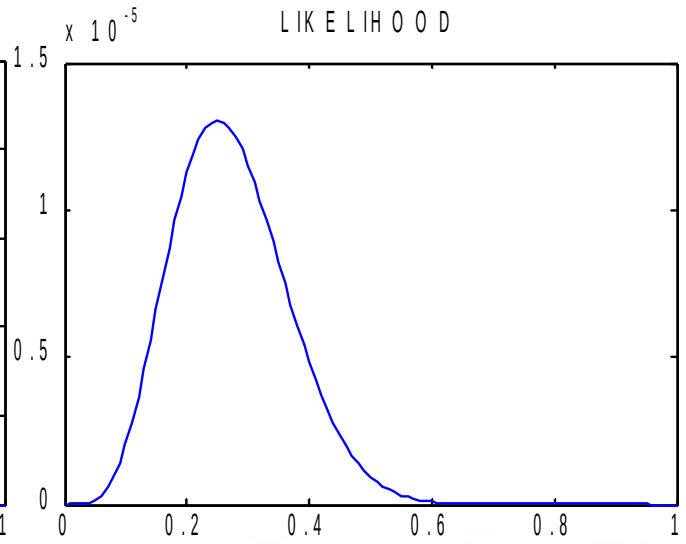
- Now let's take a coin from Bank of America. This prior is narrower (a.k.a. “stronger”).

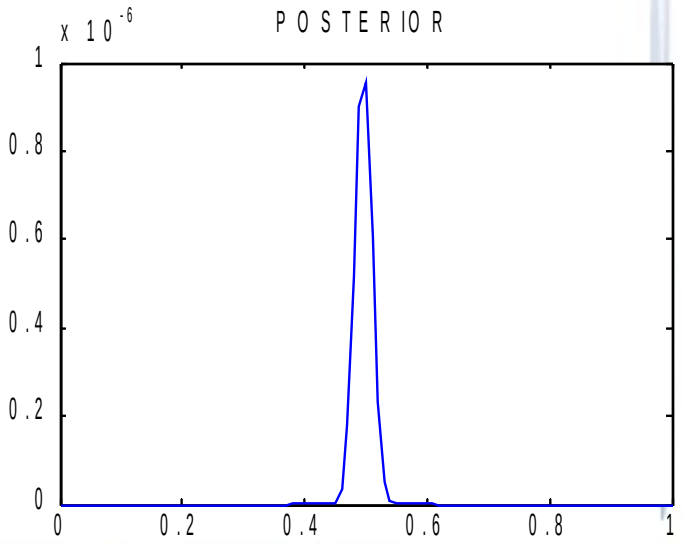
$$P(\theta|D) = P(\theta) * P(D|\theta) / P(D)$$

posterior = prior * likelihood (normalized)

- Now $\theta_{MAP} \sim= .49$

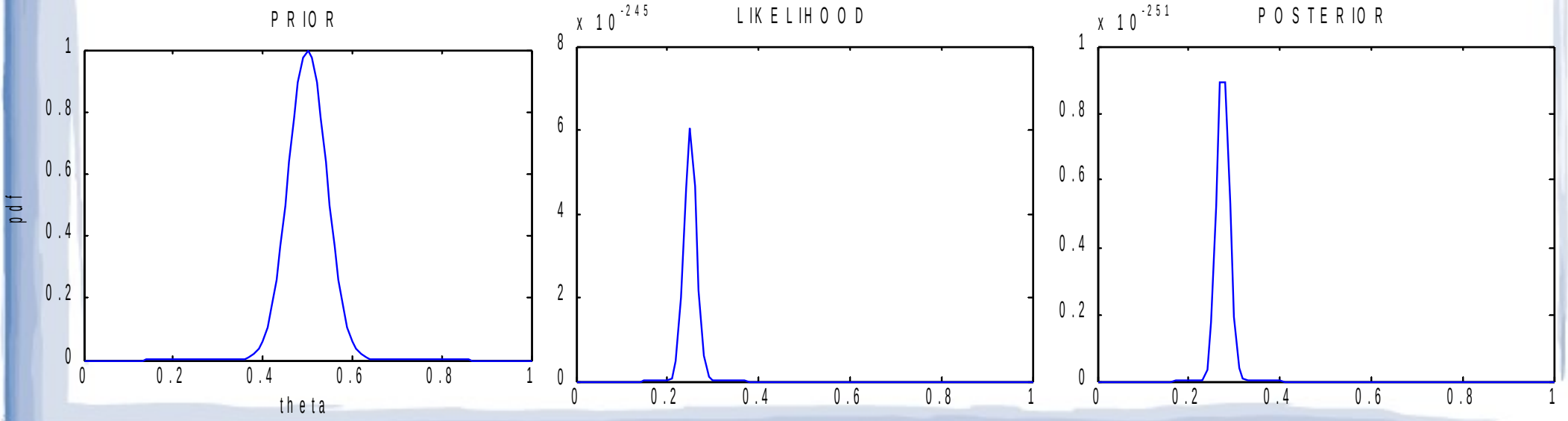






Enough data will wash away the prior

- If you have enough data and the prior is nonzero at the truth, then the prior gets washed away eventually.
- e.g. 1000 flips: 250 heads, 750 tails.
- Now $\theta_{MAP} \sim= .27$



Point estimates: MLE vs MAP

- MLE and MAP are **point estimates**.

- Mathematically, likelihood = posterior with a uniform prior (But sometimes, a uniform prior isn't a real distribution... e.g. a uniform prior over the real numbers is an **improper prior**. Many Bayesians do not like them).

- Good summary of your conclusion, unless large / multiple regions of high likelihood / probability

Newspaper headline:

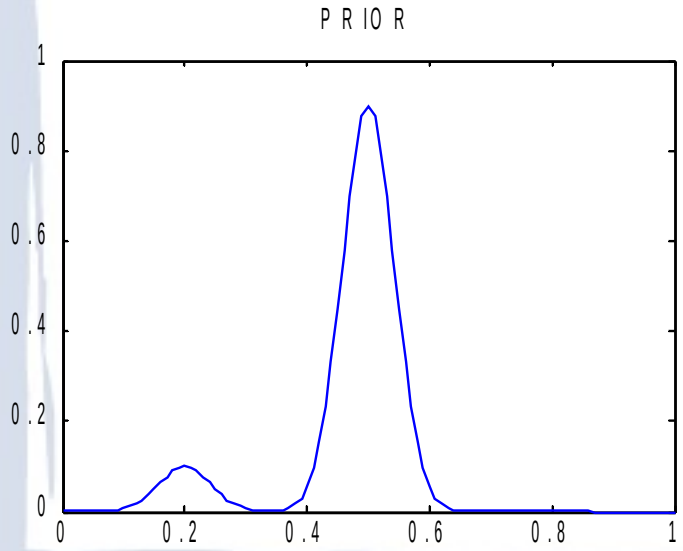
"Santa Fe Flooded with Cheater coins! Only 1/5 chance of heads!!"

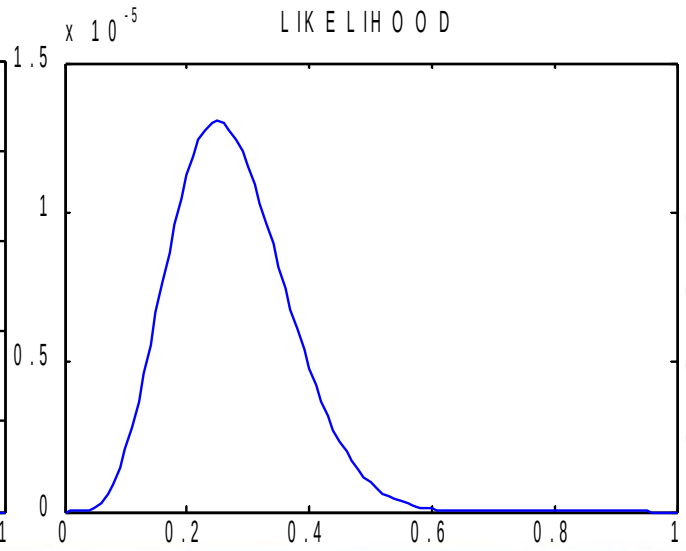
- Relevant knowledge => we need to modify our prior!

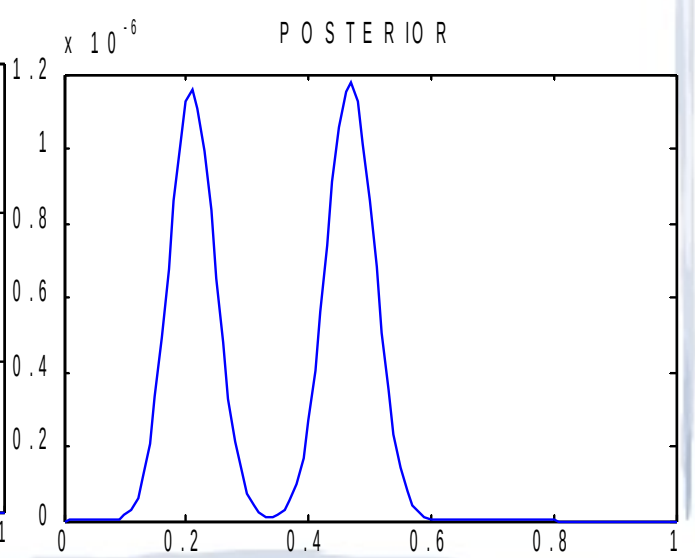
Assume 1/10 are cheater coins, then we can use this prior.

- But now the posterior has 2 roughly equal modes!

MAP is no good here...



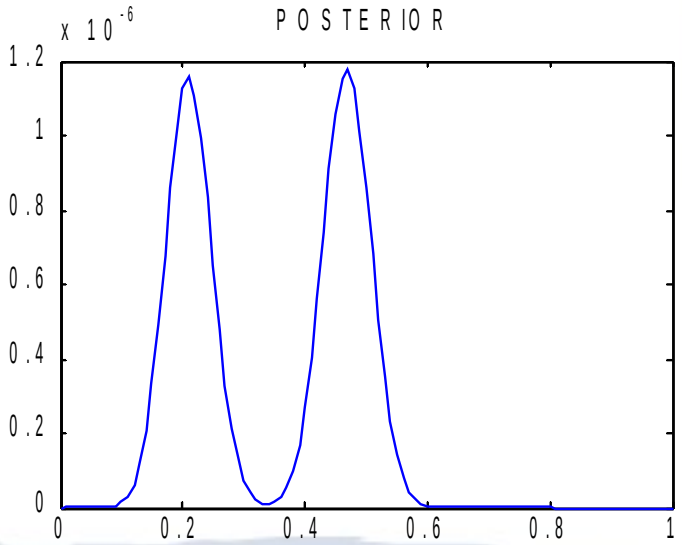




Bayesian Model Averaging

- Instead of just taking the MAP, we can be fully Bayesian, by embracing the uncertainty in the posterior.

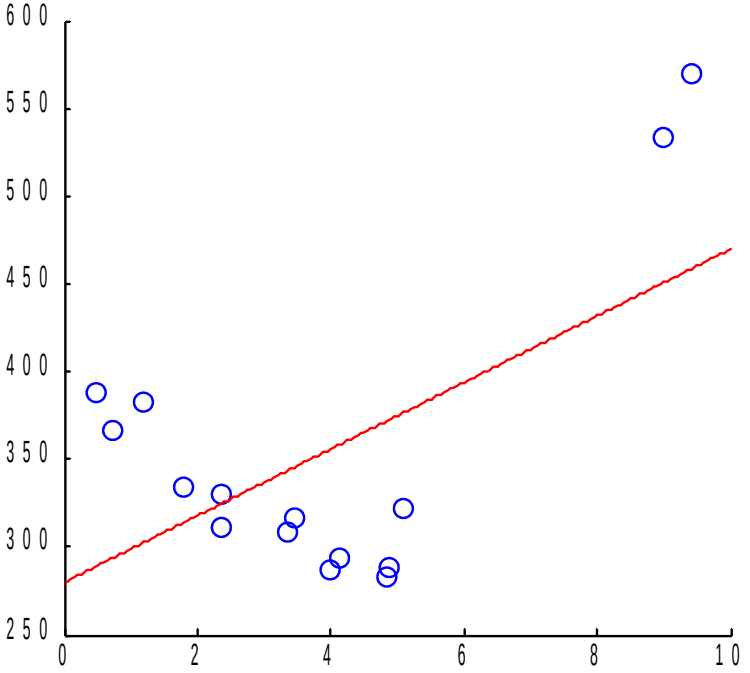
- We average our quantity of interest according to the posterior.

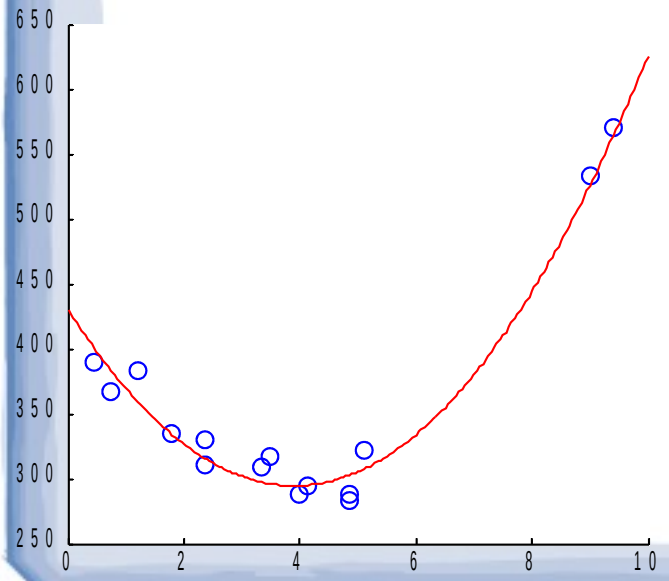


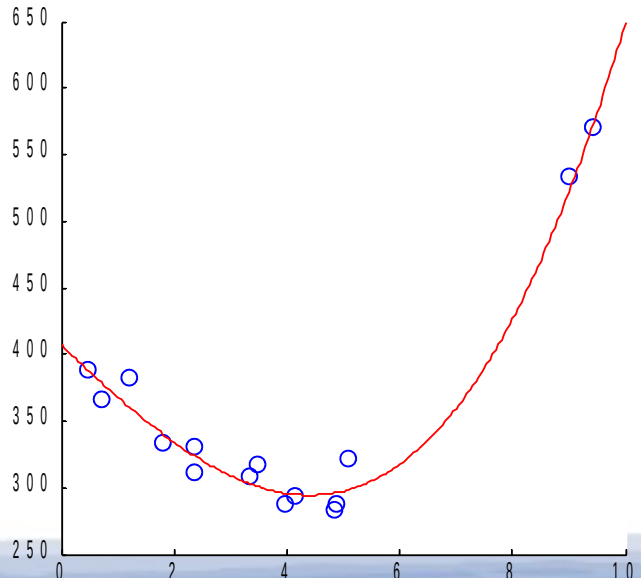
Polynomial regression

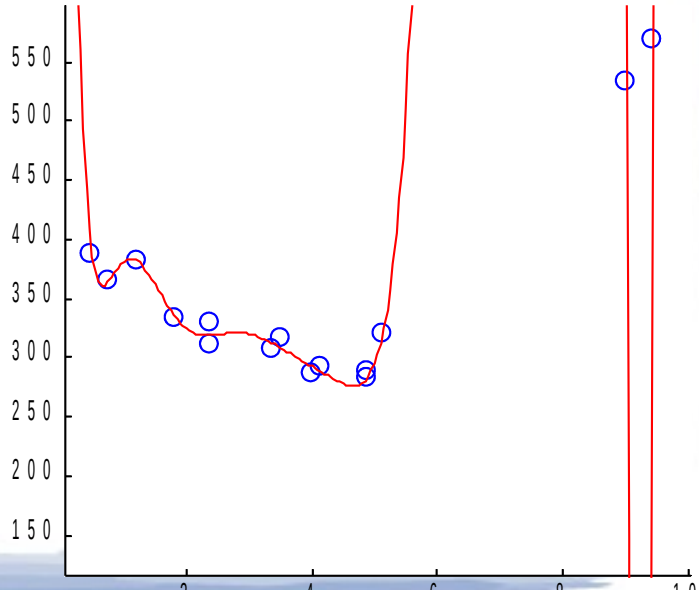
- Suppose you see some data like this:
- Linear regression is clearly a bad model.
- Quadratic looks much better.
- Cubic fits even better!
- 10^{th} degree even better!!
- The **training error** always decreases when you add more features...

- So WHERE DO YOU STOP???







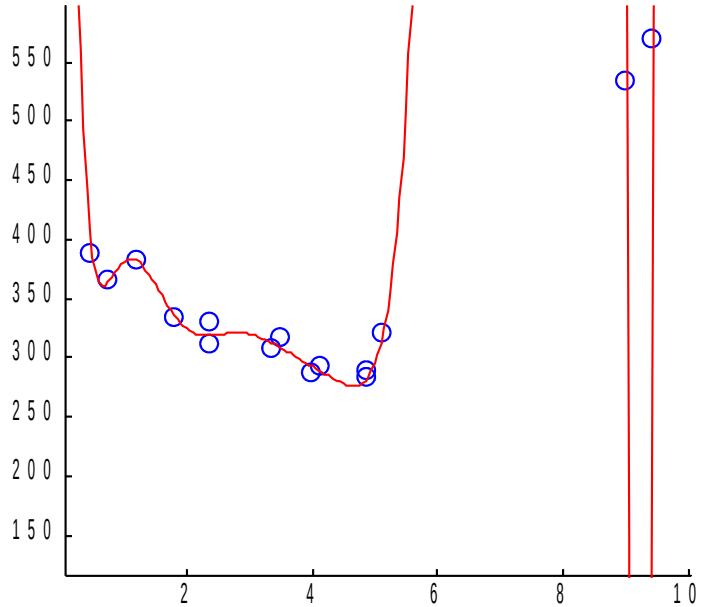


Overfitting

• For a model with parameters up to the 10^{th} degree, the MLE is the curve shown.

This curve is massively overfit.

Q: How can we prevent overfitting?



• **Cross-validation:** divide the data into training and test sets.

• **Regularization:** we maximize a **penalized likelihood**
We penalize complex models and/or big coefficients (this usually has a Bayesian interpretation, in terms of **Occam priors**)

• Choosing the degree of the polynomial is an instance of **model selection**.

Penalized Likelihood

- Can be interpreted as Bayesian inference, i.e. prior on the coefficients

- L2 penalty (Euclidean norm):
  - Shrink towards zero
  - Corresponds to Gaussian prior

- L1 penalty (Manhattan norm):
  - Sparse vector (tends to make some entries exactly zero)
  - Corresponds to Laplace prior

When and why be Bayesian

- Prior knowledge is strong relative to data
  - e.g. more features than data points:
    - gene microarray
    - computer vision

- Integrate related problems, to form a unified theory:
  - Structured Output Classification: e.g. classify nodes in a network, but each label is informative of its neighbors' labels.
  - See Hierarchical Bayes (at the end)

- *“at least, we're honest about our prejudices.”*

- Can easily make meta-models, which could automatically change your model as the data suggests it.

- Two-envelopes paradox: solved by having a proper prior.

Why NOT be Bayesian

- Priors are SUBJECTIVE!!!
  - **Probability elicitation** usually gives inconsistent results, even with experts who know their probability.

- Bayesians have to do lots of integrals (since they marginalize over so many parameters).

- Frequentists (usually) have ways of accomplishing the same things as Bayesians.

- **Beware of Bayesians:** they often choose **conjugate priors**, only because they make Bayesian updating analytically tractable.

- The frequentist/Bayesian debate will never end...

Supervised Learning

- **Problem:** reverse-engineer a function based on <x, f(x)> pairs.

Given input ***x*** and output ***y***, and some (stochastic) function f:

$$y = f(x)$$

Figure out *f* by observing the *x*'s and *y*'s

- **Discrete output:** classification
- **Continuous output:** regression
- In unsupervised learning, there is no **x**.

Probabilistic Graphical Models

- Help handle multivariate data, when you know (or can test) certain **conditional independences**.

- They represent families of joint distributions (with the parameters, they represent specific distributions)

- **Intuitive:** e.g. how much you eat today depends on how hungry you are, but only *indirectly* on how much you ate yesterday.

- Directed PGMs ~= “Bayesian Networks” (which are NOT Bayesian)

- Markov Blanket for DGMs

Probabilistic Graphical Models

- HMM, 2nd order HMM
- d-separation, the Bayes-Ball algorithm
- Explaining away: sprinkler → grass wet ← rain
- Undirected Models:
  - Markov Blanket for UGMs.
  - potential functions, e.g. Ising model

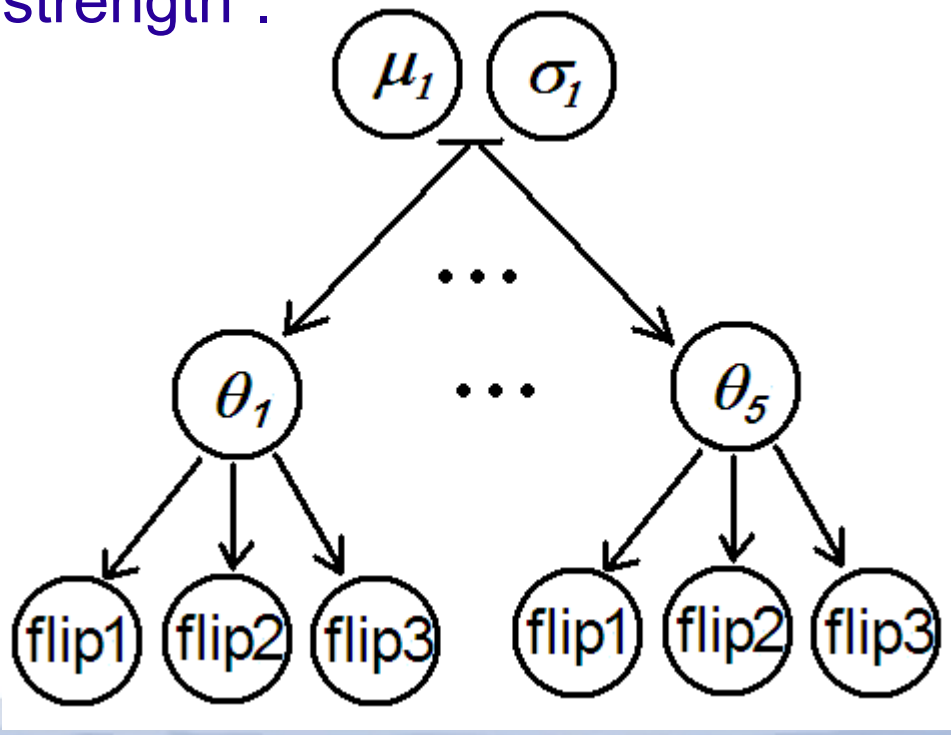
Causal Networks

- DGMs with a causal interpretation
- These models predict what would happen if you performed a given intervention.
- Can be discovered from observational data under certain assumptions (Pearl 1993; Spirtes 2002; Shimizu 2006)

Hierarchical Bayes and shrinkage

- Imagine grabbing 5 coins from some unknown source, and flipping each 3 times...

- Each individual coin has its own θ, sampled from a Gaussian
- This structure will **shrink** the θ's towards their mean (the smaller the σ, the more you shrink). Known as "borrowing statistical strength".



- μ, σ are known as **hyper-parameters**.

- You can always add more layers of uncertainty and structure, *ad infinitum*...

My view of Machine Learning