An Intro To Bayesian ML (Part 1)

Prof. Chris Jermaine
cmj4@cs.rice.edu

Before We Begin

- In A8, implement a ML algorithm to extract “topics” from text
- This is not a ML class... so to be fair
  - — Will try to specify algorithm so precisely in the next few lectures
  - — That you can implement it without really understanding what is going on
- That said...
  - — It would be a shame if this is what happend!
  - — So will spend considerable time trying to explain what’s going on
  - — And I hope it’s gonna make sense!
- So sit back, enjoy, and hopefully you’ll learn something!

Modern Machine Learning

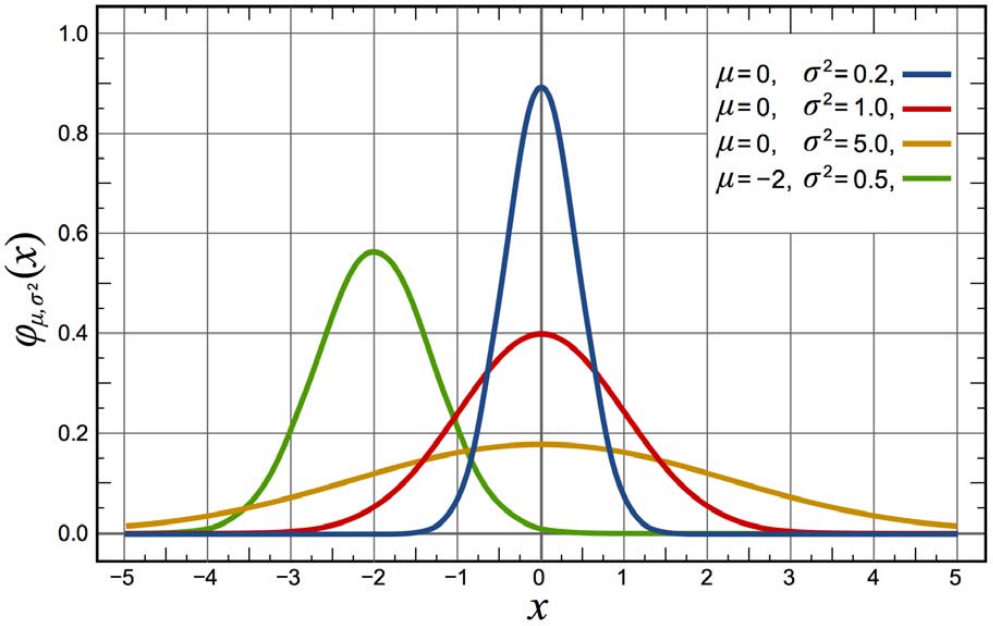
- Is a whole lot like applied statistics
- In fact, I'd argue that CS in general is becoming more statistical
  - — That is, based on observation and inference
  - — ...and a bit less logic-based
- Ask Google, Facebook, LinkedIn how important stats is!

The Bayesian Approach

- One branch of machine learning utilizes the “Bayesian” approach
- Say our goal was to give an exam to a few students
  - — Then use their performance to figure out what the class average will be
  - — Want to figure out if exam is too hard or too easy
- How would a Bayesian do this?

So How Would a Bayesian Do This?

- First imagine a stochastic "generative process" for the data
  - — For the *i*th student, we might imagine $\text{score}_i \sim \text{Normal}(\text{mu}, \text{sigma}^2)$

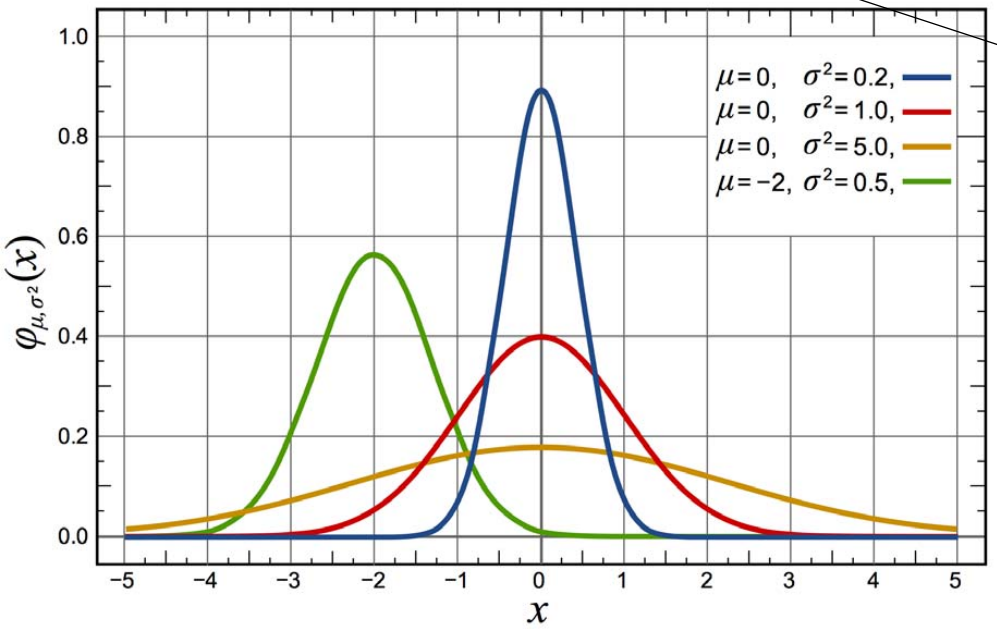


  - — Why normal? Models typical "bell shaped" data
  - — Assume mu, sigma^2 are also generated by sampling from RVs, and are unknown

So How Would a Bayesian Do This?

ultimate goal is to guess the value of mu

- First imagine a stochastic "generative process" for the data
  - — For the *i*th student, we might imagine $score_i \sim Normal(mu, sigma^2)$

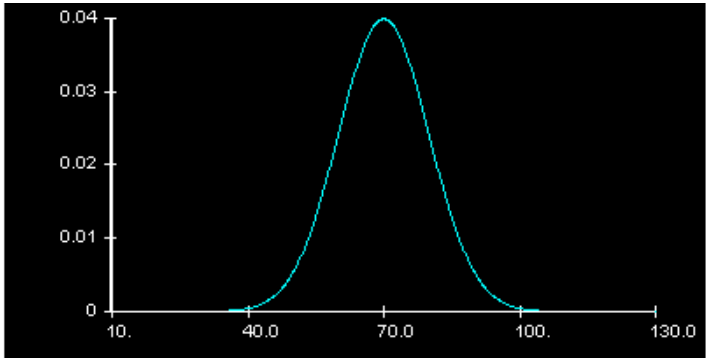


Means "is sampled from" a RV with a normal dist.

  - — Why normal? Models typical "bell shaped" data
  - — Assume mu, $sigma^2$ are also generated by sampling from RVs, and are unknown

So How Would a Bayesian Do This?

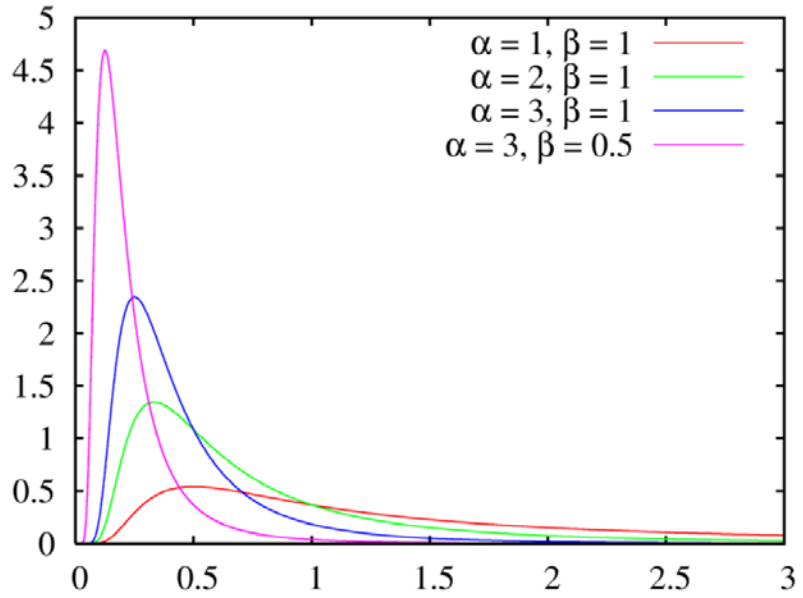
- How is the mean mu generated?
  - — We imagine mu ~ Normal (70, 100)



  - — Why Normal (70, 100)? Allows for all possible, reasonable exam averages

So How Would a Bayesian Do This?

- How is the variance (spread) sigma^2 generated?
  - — We imagine sigma^2 ~ InverseGamma (1, 1)



  - — Why InverseGamma (1, 1)? Allows a really large range of positive sigma^2 vals

Thus, Our Generative Process Is

— Step 1: $\text{sigma}^2 \sim \text{InverseGamma}\ (1, 1)$

— Step 2: $\text{mu} \sim \text{Normal}\ (75, 100)$

— Step 3: for each i, $\text{score}_i \sim \text{Normal}\ (\text{mu}, \text{sigma}^2)$

Why Have a Generative Process?

- Given gen process, can measure how likely a given pair is
  - Specifically, $p(\mu, \sigma^2) = IG(\sigma^2 \mid 1, 1) \times Normal(\mu \mid 75, 100)$
  - So given any mu, σ^2 combo, we can say how likely we think it is
- This is our "prior" on mean and variance

Note "times" cause the two values are indep.

Bayesian Inference

- To a Bayesian, “inference” is then the process of updating prior expectations after seeing the data
- After we see the data: test scores

  — $p(\text{mu}, \text{sigma}^2 | \text{data}) = \frac{p(\text{data} | \text{mu}, \text{sigma}^2) \times p(\text{mu}, \text{sigma}^2)}{p(\text{data})}$

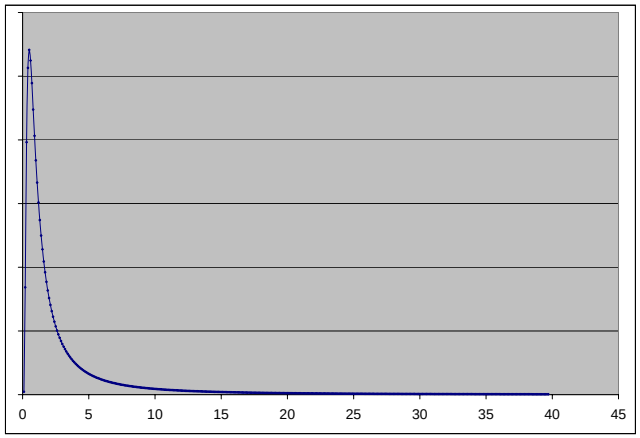
  — This is equation follows directly from “Bayes’ Theorem”

  — Note: this is also a distribution

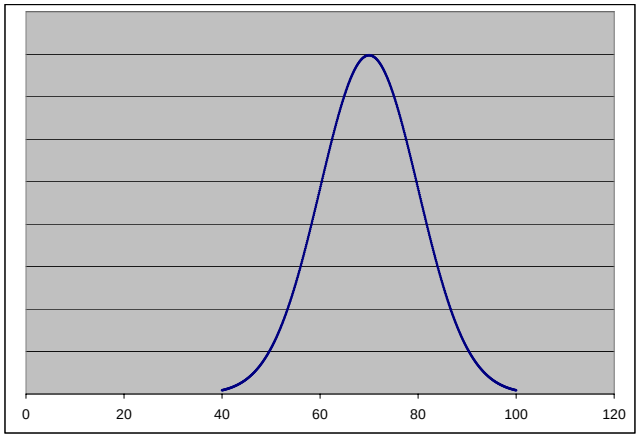
  — Known as a “posterior” distribution

Pictorially

- You have a prior distribution on sigma^2 and mu:



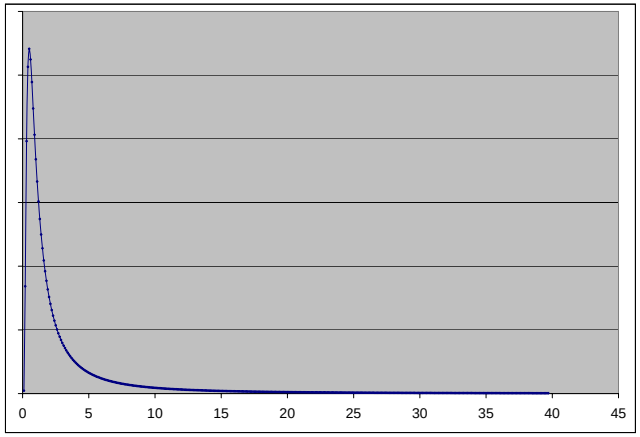
sigma^2
(variance of test scores)



mu
(average of test scores)

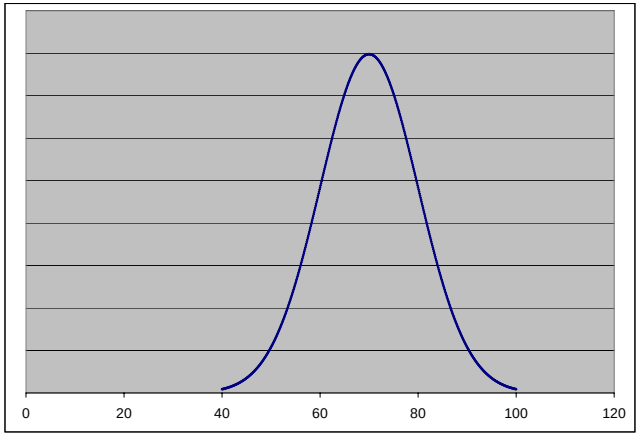
Pictorially

- You see some test scores



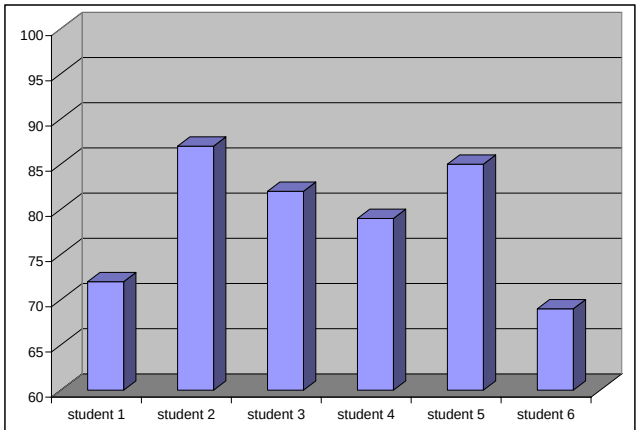
sigma^2

(variance of



mu

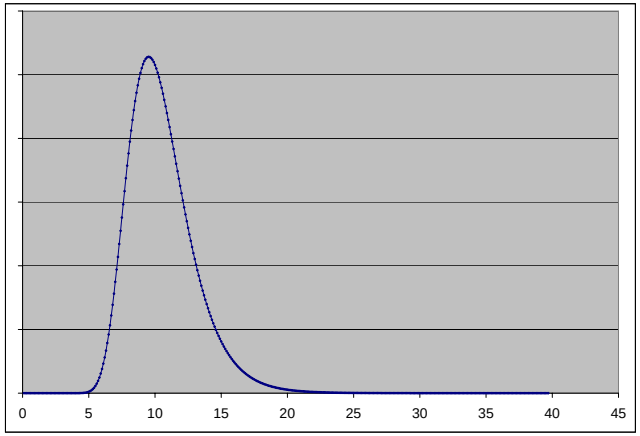
rage of test scores)



seem to average in the high 70's

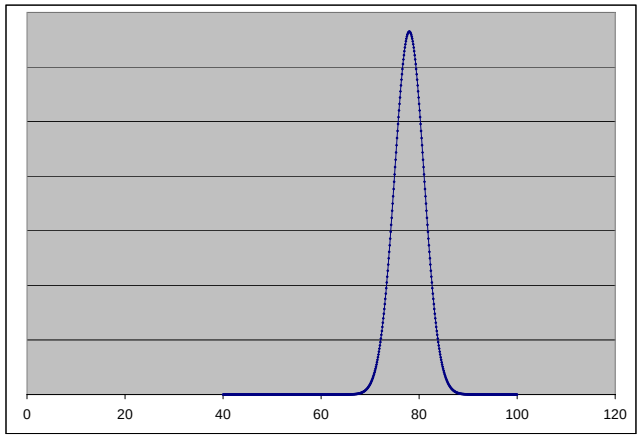
Pictorially

- Then use Bayes' to get a posterior distribution on sigma^2 and mu:



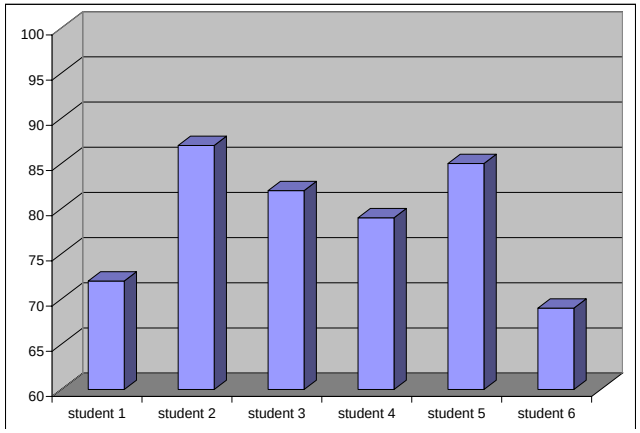
sigma^2

(variance of



mu

erage of test scores)



much narrower
after seeing data!

That's the Bayesian Approach

- Come up with a generative process
- Which includes prior distributions on the quantities like to est
  - In our example, the variance and especially the mean
- See some data
- Use Bayes' Theorem and data to "update" the priors
  - This gives you a posterior dist
  - The posterior contains your estimate

Questions?