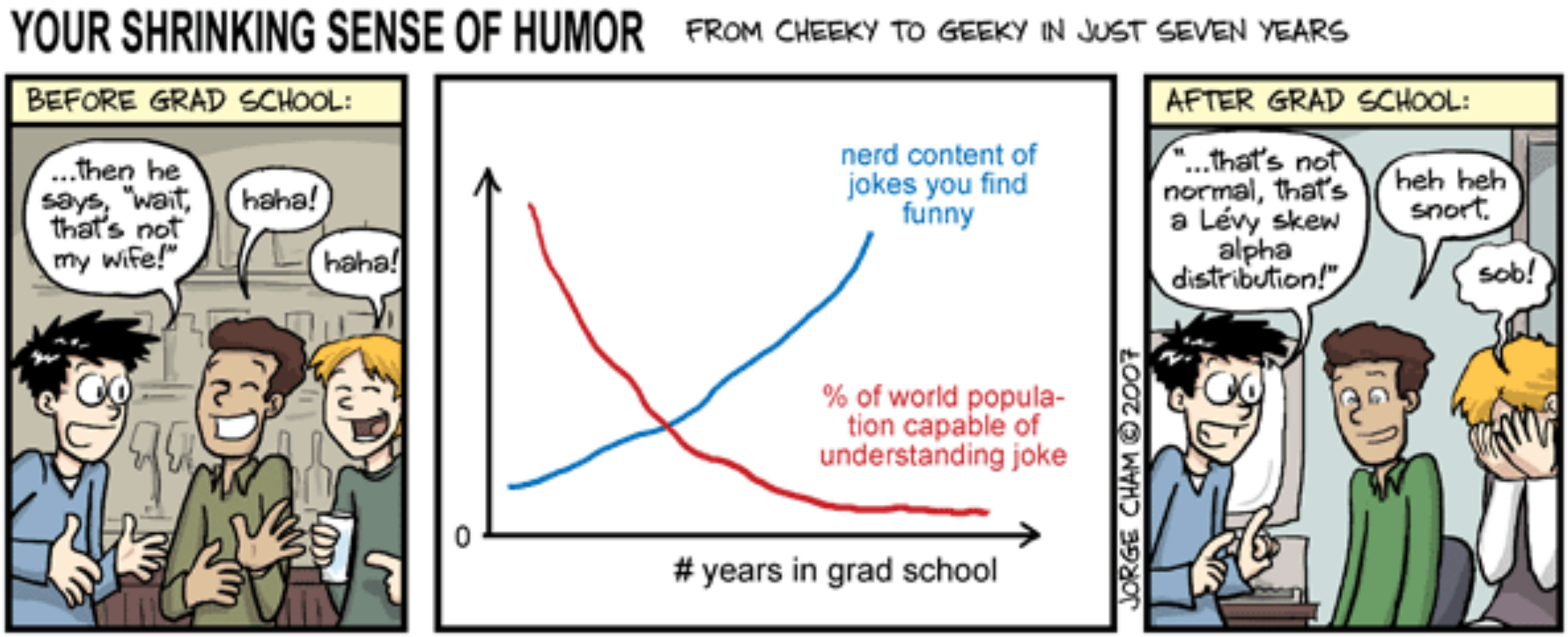
COMPUTER SCIENCE
INDIANA UNIVERSITY
School of Informatics and Computing
Bloomington

Intro to prediction problems and Linear regression



Reminders

- Assignment #2 is due this week on Wednesday
- Next readings: Chapters 3 and 4
  - Chapter 3: Introduction to machine learning
  - Chapter 4: Linear regression
- Calculus and derivatives refresher?
  - if there are steps you do not follow, interrupt me

Assignment 1 clarifications

- Question 8: Borel algebra on [0,1]
  - made up of intervals of form (a,b) (i.e., open), [a,b] (i.e., closed) and half-open (a,b] and [a, b)
  - has non-contiguous intervals: (a,b) union (c,d)

- Any other clarifications needed?

- Note: the AIs will mark the assignments, so do go to their office hours

Thought questions

- Interesting thoughts!

- Submit questions from different sections (and label the section in which you had the thought question)
  - if it is a more general thought after the readings, no need, but make the thought questions different (and about different ideas/issues)

- Avoid clarification/confusion questions
  - e.g. I still do not get what a model is
  - You should understand the material and give an insightful next question about it; if you do not understand the material, try to understand it first
  - Best strategy: first explain what you understood, then give a follow-up thought question

Example of thought question

- **“Bad” thought question:** I still do not understand what a model is. Are they distributions? Can they be other things?
  - No one asked this question, per se

- **Alternative:** The notion of a model appears to be somewhat imprecise. We have used distributions as a model of our data, with parameters to those distributions representing the model. But, can other thing be models? For example, is plotting the data points and understanding its behavior considered a “model” of the data? What other kinds of models are there?
  - First showed that understood how we have been describing models
  - Then showed follow-up thought about what the term “model” could really mean

Example of thought question

- **"Bad" thought question:** What are the necessary assumptions, if any, in applying MAP estimation for model selection?

- **Alternative:** We seem to be able to apply MAP to a likelihood, prior combination, simply by taking the log to separate across points and to separate the likelihood and prior. To make the likelihood separate into individual probabilities, we have assumed that the data are iid. Is this a necessary assumption? Are there other implicit assumptions being made? Are these assumptions only needed to make the computation feasible? In general, can we specify the MAP optimization for any likelihood and prior, even if we cannot compute the solution?

  - First showed that understood how we have been using MAP
  - Then showed follow-up thought about generally using MAP

Why this focus on thought questions?

- Whether academia or industry, specifying projects involves understanding what exists, and proposing the “next” thing

- This includes identifying
  - current assumptions/beliefs that could be challenged
  - gaps in current approaches (practical/theoretical)
  - limitations, so can keep those limitations in mind for the solution
  - novel ways forward, given the current solutions/understanding

- Secondary reason: many people want to be data scientists; the way to set yourself apart is to have a deeper understanding of fundamentals and novel insights

Exercise: mixture model

- Let's explicitly go through the steps of computing the expected value and the variance

- Recall: expectation of a function of a random variable

$$E\left[f(X)\right] = \begin{cases} \sum_{x\in\Omega_X} f(x)p_X(x) & X : \text{discrete} \\ \\ \int_{\Omega_X} f(x)p_X(x)dx & X : \text{continuous} \end{cases}$$

- Mixture model: $p(x) = \sum_{i=1}^{m} w_i p_i(x),$

Exercise: mixture model

$$E\left[f(X)\right] = \int_{-\infty}^{+\infty} f(x) p_X(x) dx$$
$$= \int_{-\infty}^{+\infty} f(x) \sum_{i=1}^{m} w_i p_{X_i}(x) dx$$
$$= \sum_{i=1}^{m} w_i \int_{-\infty}^{+\infty} f(x) p_{X_i}(x) dx$$
$$= \sum_{i=1}^{m} w_i E[f(X_i)].$$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x),$$

We can now apply this formula to obtain the mean, when $f(x) = x$ and the variance, when $f(x) = (x - E[X])^2$, of the random variable X as

$$E[X] = \sum_{i=1}^{m} w_i E[X_i],$$

and

$$V[X] = \sum_{i=1}^{m} w_i V[X_i] + \sum_{i=1}^{m} w_i \left(E[X_i] - E[X]\right)^2,$$

Summary

- We will learn parameters to distribution models
  - this includes conditional distributions p(y | x, w) and general distributions p(x | theta)
- We will formalize the problem using maximum likelihood, MAP
- In maximum likelihood, we find optimal parameters theta for p(D | theta) (i.e., theta that makes data most likely)
- In MAP, also have expert defined prior over parameters p(theta), and optimize p(D | theta) p(theta) (i.e., theta that makes data most likely, but also satisfies prior as much as possible to balance the two)
- For both, we define optimization argmax, and use derivatives

Back to prediction

- Discussed notation and some goals for prediction
- Introduced different types of features and targets
- We will now formalize our measure of success for both regression and classification

Optimal prediction

- We want to learn a prediction function
- We will need to define cost/error of a prediction
  - Cost function $c : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \infty)$
  - For true target y, get cost $c(\hat{y}, y)$
- We will see that modeling p(y l x) is useful for this task
- Want to find predictor that minimizes the expected cost
  - could choose other metrics, such as minimize number of costs that are very large or minimize the maximum cost

Whiteboard

- Expected cost
- Bayes optimal models