Accelerated Natural Language Processing
Lecture 4
Models and probability estimation

Sharon Goldwater

23 September 2019

School of informatics

A famous quote

> It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.
> Noam Chomsky, 1969

A famous quote

> It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.
> Noam Chomsky, 1969

- "useless": To everyone? To linguists?
- "known interpretation": What are possible interpretations?

Today's lecture

- What do we mean by the "probability of a sentence" and what is it good for?
- What is probability estimation? What does it require?
- What is a generative model and what are model parameters?
- What is maximum-likelihood estimation and how do I compute likelihood?

Intuitive interpretation

- "Probability of a sentence" = how likely is it to occur in natural language
  - Consider only a specific language (English)
  - Not including meta-language (e.g. linguistic discussion)

P(She studies morphosyntax) > P(She studies more faux syntax)

Automatic speech recognition

Sentence probabilities (**language model**) help decide between similar-sounding options.

speech input

↓ (Acoustic model)

possible outputs
- She studies morphosyntax
- She studies more faux syntax
- She's studies morph or syntax
- ...

↓ (Language model)

best-guess output: She studies morphosyntax

Machine translation

Sentence probabilities help decide word choice and word order.

non-English input

↓ (Translation model)

possible outputs
- She is going home
- She is going house
- She is traveling to home
- To home she is going
- ...

↓ (Language model)

best-guess output: She is going home

So, not "entirely useless"...

- Sentence probabilities are clearly useful for language engineering [this course].
- Given time, I could argue why they're also useful in linguistic science (e.g., psycholinguistics). But that's another course...

But, what about zero probability sentences?

the Archaeopteryx winged jaggedly amidst foliage
vs
jaggedly trees the on flew

- Neither has ever occurred before.
  ⇒ both have zero probability.
- But one is grammatical (and meaningful), the other not.
  ⇒ "Sentence probability" is useless as a measure of grammaticality.

The logical flaw

- "Probability of a sentence" = how likely is it to occur in natural language.
- Is the following statement true?

  Sentence has never occurred ⇒ sentence has zero probability
- More generally, is this one?

  Event has never occurred ⇒ event has zero probability

Events that have never occurred

- Each of these events has never occurred:

  My hair turns blue
  I injure myself in a skiing accident
  I travel to Finland
- Yet, they clearly have different (and non-zero!) probabilities.

Events that have never occurred

- Each of these events has never occurred:

  My hair turns blue
  I injure myself in a skiing accident
  I travel to Finland
- Yet, they clearly have differing (and non-zero!) probabilities.
- Most sentences (and events) have never occurred.
  - This doesn't make their probabilities zero (or meaningless), but
  - it does make **estimating** their probabilities trickier.

Probability theory vs estimation

- Probability theory can solve problems like:
  - I have a jar with 6 blue marbles and 4 red ones.
  - If I choose a marble uniformly at random, what's the probability it's red?
- But what about:
  - I have a jar of marbles.
  - I repeatedly choose a marble uniformly at random and then replace it before choosing again.
  - In ten draws, I get 6 blue marbles and 4 red ones.
  - On the next draw, what's the probability I get a red marble?
- The latter also requires estimation theory.

Example: weather forecasting

What is the probability that it will rain tomorrow?

- To answer this question, we need
  - data: measurements of relevant info (e.g., humidity, wind speed/direction, temperature).
  - model: equations/procedures to estimate the probability using the data.
- In fact, to build the model, we will need data (including *outcomes*) from previous situations as well.

Example: weather forecasting

What is the probability that it will rain tomorrow?

- To answer this question, we need
  - data: measurements of relevant info (e.g., humidity, wind speed/direction, temperature).
  - model: equations/procedures to estimate the probability using the data.
- In fact, to build the model, we will need data (including *outcomes*) from previous situations as well.
- Note that we will never know the "true" probability of rain $P(\text{rain})$, only our estimated probability $\hat{P}(\text{rain})$.

Example: language model

What is the probability of sentence $\vec{w} = w_1 \dots w_n$?

- To answer this question, we need
  - data: words $w_1 \dots w_n$, plus a large corpus of sentences ("previous situations", or **training data**).
  - model: equations to estimate the probability using the data.
- Different models will yield different estimates, even with the same data.
- Deep question: what model/estimation method do humans use?

How to get better probability estimates

Better estimates definitely help in language technology. How to improve them?

- **More training data.** Limited by time, money. (Varies a lot!)
- **Better model.** Limited by scientific and mathematical knowledge, computational resources
- **Better estimation method.** Limited by mathematical knowledge, computational resources

We will return to the question of how to know if estimates are "better".

Notation

- When the distinction is important, will use
  - $P(\vec{w})$ for *true* probabilities
  - $\hat{P}(\vec{w})$ for *estimated* probabilities
  - $P_{\mathrm{E}}(\vec{w})$ for estimated probabilities using a particular estimation method E.
- But since we almost always mean estimated probabilities, may get lazy later and use $P(\vec{w})$ for those too.

Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is $\hat{P}(\mathrm{T})$?

Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is $\hat{P}(\mathrm{T})$?

- **A:** $\hat{P}(T) = 0.5$
- **B:** $\hat{P}(T) = 0.7$
- **C:** Neither of the above
- **D:** I don't know

Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is $\hat{P}(\text{T})$?

- **Model 1:** Coin is fair. Then, $\hat{P}(T) = 0.5$

Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is $\hat{P}(\text{T})$?

- **Model 1:** Coin is fair. Then, $\hat{P}(T) = 0.5$
- **Model 2:** Coin is not fair.[1] Then, $\hat{P}(T) = 0.7$ (why?)

[1] Technically, the physical process of flipping a coin means that it's not really possible to have a biased coin flip. To see a bias, we'd actually need to *spin* the coin vertically and wait for it to tip over. See `https://www.stat.berkeley.edu/~nolan/Papers/dice.pdf` for an interesting discussion of this and other coin flipping issues.

Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is $\hat{P}(\text{T})$?

- **Model 1:** Coin is fair. Then, $\hat{P}(T) = 0.5$
- **Model 2:** Coin is not fair. Then, $\hat{P}(T) = 0.7$ (why?)
- **Model 3:** Two coins, one fair and one not; choose one at random to flip 10 times. Then, $0.5 < \hat{P}(T) < 0.7$.

Example: estimation for coins

I flip a coin 10 times, getting 7T, 3H. What is $\hat{P}(\text{T})$?

- **Model 1:** Coin is fair. Then, $\hat{P}(T) = 0.5$
- **Model 2:** Coin is not fair. Then, $\hat{P}(T) = 0.7$ (why?)
- **Model 3:** Two coins, one fair and one not; choose one at random to flip 10 times. Then, $0.5 < \hat{P}(T) < 0.7$.

Each is a **generative model**: a probabilistic process that describes how the data were generated.

Defining a model

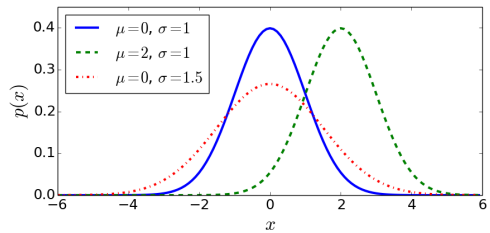
Usually, two choices in defining a model:

- **Structure** (or **form**) of the model: the form of the equations, usually determined by knowledge about the problem.
- **Parameters** of the model: specific values in the equations that are usually determined using the training data.

Example: height of 30-yr-old females

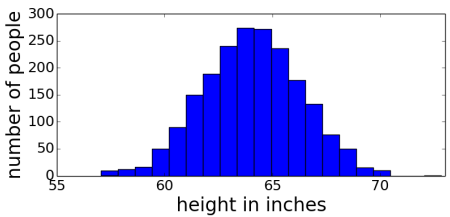
Assume the form of a **normal distribution** (or **Gaussian**), with parameters (μ, σ):

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$



Example: height of 30-yr-old females

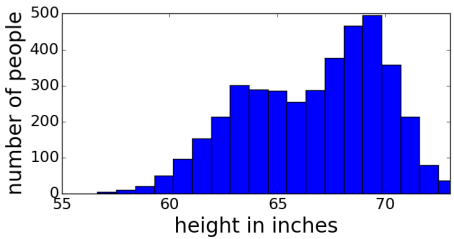
Collect data to determine values of μ, σ that fit this particular dataset.



I could then make good predictions about the likely height of the next 30-yr-old female I meet.

Example: height of 30-yr-old females

What if our data looked like this?



Model criticism

- Sometimes using an incorrect model structure can still give useful results. (We'll see examples.)
- But sometimes we might need to revise the model if the data don't seem to match the model assumptions.

All models are approximations. How good the approximation needs to be depends on what we are trying to do with it.

The true model

The **true** generative model for the second dataset was actually:

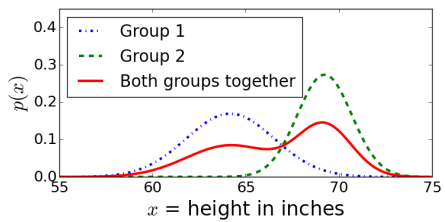
Assume two groups, each with a Gaussian distribution.
For each data point,
1. Choose which group this point belongs to.
2. *Conditioned on* the group, choose height value from that group's distribution.

Question: how many parameters does this model have?

Mixture model

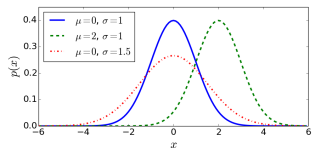
This model is a **mixture** of two Gaussians, and has **five** parameters:

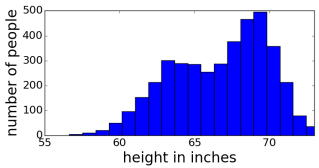
- The **mixing weight**: probability of choosing group 1 or group 2 (in this case, 0.5).
- μ and σ for each of the two Gaussian distributions.



Mixture model

If I use the original model structure (single Gaussian), no estimate of model parameters will lead to accurate predictions.





Example: M&M colors

What is the proportion of each color of M&M?

- Assume a **discrete distribution** with parameters θ.
  - θ is a vector! That is, $\theta = (\theta_{\text{R}}, \theta_{\text{O}}, \theta_{\text{Y}}, \theta_{\text{G}}, \theta_{\text{BL}}, \theta_{\text{BR}})$.
  - For discrete distribution, params ARE the probabilities, e.g., $P(\text{red}) = \theta_{\text{R}}$.
  - Note: if there are six colors, there are really only **five** parameters. (why?)

Example: M&M colors

What is the proportion of each color of M&M?

- Assume a **discrete distribution** with parameters θ.
- In 48 packages, I find[2] 2620 M&Ms, as follows:

Red	Orange	Yellow	Green	Blue	Brown
372	544	369	483	481	371

- How to estimate θ from this data?

[2] Actually, data from: https://joshmadison.com/2007/12/02/mms-color-distribution-analysis/

Relative frequency estimation

- Intuitive way to estimate discrete probabilities:

$$P_{\text{RF}}(x) = \frac{C(x)}{N}$$

where $C(x)$ is the count of x in a large dataset, and $N = \sum_{x'} C(x')$ is the total number of items in the dataset.

- M&M example: $P_{\text{RF}}(\text{red}) = \hat{\theta}_{\text{R}} = \frac{372}{2620} = .142$
- Or, could estimate probability of word w from a large corpus.
- Can we justify this mathematically?

Relative frequency estimation

As the number of observations approaches infinity, relative frequency estimate converges to the true probability. In practical terms,

- If our counts are **large**, estimates are fairly accurate.
  150 red M&Ms of out 1000:
  $P_{\text{RF}}(red) = .15$ and $P(red)$ not likely to be .1 or .2.
- If our counts are **small**, estimates are not so accurate.
  3 red M&Ms of out 20:
  $P_{\text{RF}}(red) = .15$ but $P(red)$ could easily be .1 or .2.

(It's really the size of the numerator that matters, as we'll see later.)

Maximum-likelihood estimation

RF estimation is also called **maximum-likelihood estimation (MLE)**.

- The **likelihood** is the probability of the observed data d under some particular model with parameters θ: that is, $P(d|\theta)$.
- For a fixed d, different choices of θ yield different $P(d|\theta)$.
- If we choose θ using relative frequencies, we get the maximum possible value for $P(d|\theta)$: the maximum likelihood.

Likelihood example

- For a fixed dataset, the likelihood depends on the model we use.
- Our coin example: $\theta = (\theta_{\mathrm{H}}, \theta_{\mathrm{T}})$. Suppose $d =$ HTTTHTHTTT.
- **Model 1:** Assume coin is fair, so $\hat{\theta} = (0.5, 0.5)$.
  - Likelihood of this model:
    $P(\mathrm{HTTTHTHTTT}|\hat{\theta}) = (0.5)^3 \cdot (0.5)^7 = 0.00097$

Likelihood example

- For a fixed dataset, the likelihood depends on the model we use.
- Our coin example: $\theta = (\theta_{\mathrm{H}}, \theta_{\mathrm{T}})$. Suppose $d =$ HTTTHTHTTT.
- **Model 1:** Assume coin is fair, so $\hat{\theta} = (0.5, 0.5)$.
  - Likelihood of this model:
    $P(\mathrm{HTTTHTHTTT}|\hat{\theta}) = (0.5)^3 \cdot (0.5)^7 = 0.00097$
- **Model 2:** Use ML estimation, so $\hat{\theta} = (0.3, 0.7)$.
  - Likelihood of this model: $(0.3)^3 \cdot (0.7)^7 = 0.00222$
- Maximum-likelihood estimate does have higher likelihood!

Questions for review:

- What do we mean by the "probability of a sentence" and what is it good for?
- What is probability estimation? What does it require?
- What is a generative model and what are model parameters?
- What is maximum-likelihood estimation and how do I compute likelihood? (more on this in the next lab.)

Where to go from here?

Next time, we'll start to discuss

- Different generative models for sentences (model structure), and the questions they can address
- Weaknesses of MLE and ways to address them (parameter estimation methods)

Exercises

1. In which of the following scenarios do we have true probabilities? In which can we only estimate probabilities?
   (a) the probability it will rain tomorrow
   (b) the probability of drawing a red ball if we choose a ball uniformly at random from a set of 3 red and 4 green balls
   (c) the probability of drawing an ace from the top of a deck of well-shuffled playing cards
   (d) the probability that the first character in an email I receive is 'a'
2. Suppose I have two 6-sided dice. One is evenly weighted, and one is unevenly weighted. I randomly pick one of the dice, then roll it and tell you the result. How many parameters are needed to fully specify a model of this generative process?

   (Hint: first think about how many parameters are needed to model each of the dice individually. Then, consider whether there are any extra parameters needed to describe the full process.)

3. Suppose I have a jar with balls of three different colours (red, green, blue). I repeatedly draw a ball out, note its colour, and then replace it in the jar. In 4 draws, I get 1 red and 3 green balls. Using maximum likelihood estimation,
   (a) what is the estimated probability of getting a red ball?
   (b) what is the estimated probability of getting a blue ball?
   (c) what is the likelihood? That is, what is the probability the model assigns to the data I observed?
   (d) what probability does the model assign to drawing the sequence red, red, blue?
4. Consider two scenarios: (a) I roll a standard six-sided die 20 times and record how many times each value comes up. (b) My friend has a device that outputs one of the numbers 1-6 at random when a button is pressed. I press the button 20 times and record how many times each value is output.

   Now I want to estimate the probability that the next outcome (dice roll or device output) will be a 1. In which scenario does it make more sense to use MLE to estimate this probability? Why?

Reminders/Announcements

- Lecture tomorrow: G.07 Meadows Lecture Theatre - Doorway 4, Medical School, Teviot
- Tutorial groups Tue/Wed/Thu: Work through problems in advance and bring questions.