STATISTICAL MODELLING

David Talbot

22nd April, 2017

Computer Science Club, St. Petersburg, Russia

STATISTICAL MODELLING

probability vs. statistics

“1/5 of the jelly beans are red. What is the probability of drawing one at random?” (Probability)

“Give me some jelly beans (at random) and I’ll tell you what proportion are red.” (Statistics)

event spaces and random variables

Probabilities are defined to sum to 1 over a set of outcomes (the event space).

For example the event space of a coin flip is $X \in \{H, T\}$ or a throw of a dice $Z \in \{1, 2, 3, 4, 5, 6\}$.

X and Z are called random variables.

joint probability

If we have two random variables X and Y, we can start to talk about the joint probability:

$$\mathbf{Pr}(X = x \cap Y = y)$$

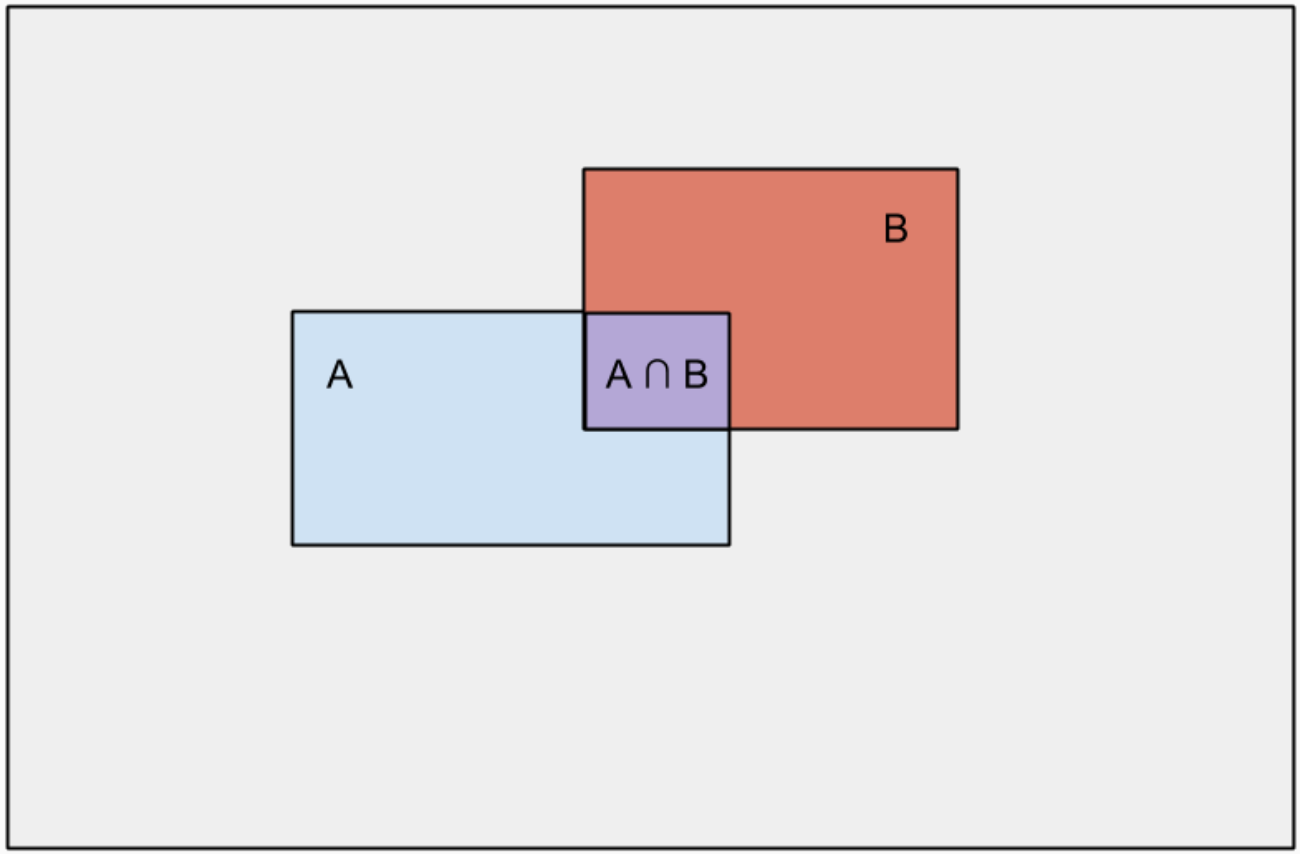
which we'll usually write as

$$\mathbf{Pr}(X = x, Y = y).$$

For example, let X be the height of a person and Y be their weight.

$$\mathbf{Pr}(X > 180\ cm, Y < 100\ kg)$$

joint probability = intersection



marginal probability

If we know $\mathbf{Pr}(X, Y)$ how can we compute $\mathbf{Pr}(X)$ or $\mathbf{Pr}(Y)$?

Name	Gender	Grade
Anna	female	5
Sarah	female	5
Maria	female	4
Dmitry	male	5
Bill	male	4
Bob	male	3

Let X be the gender of a student, i.e. $X \in \{M, F\}$
Let Y be the grade a student receives, i.e. $Y \in \{1, 2, 3, 4, 5\}$

Compute $\mathbf{Pr}(X = \mathit{male}, Y = 3)$, $\mathbf{Pr}(X = \mathit{female})$ and $\mathbf{Pr}(Y = 5)$.

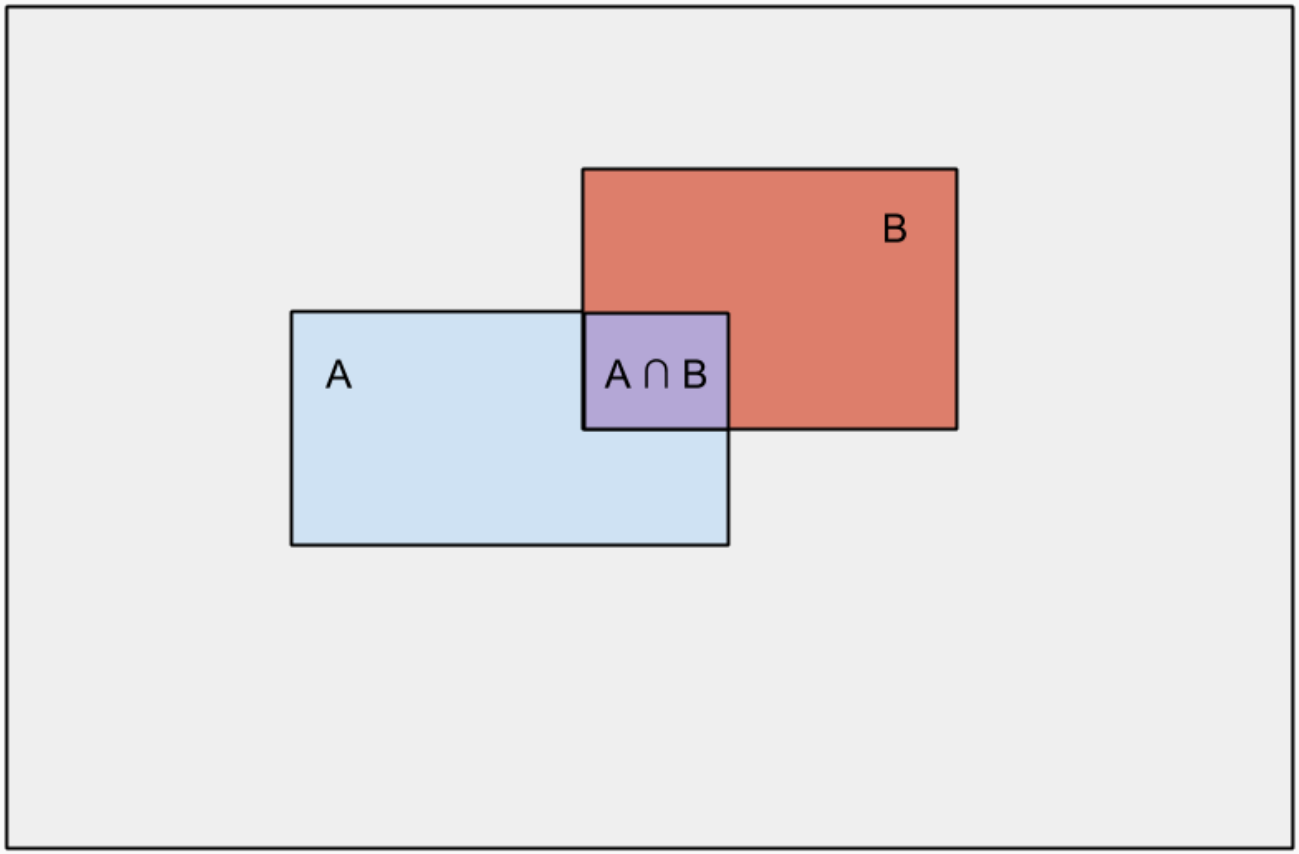
conditional probability

What's the probability that if $X = x$ then some other random variable $Y = y$?

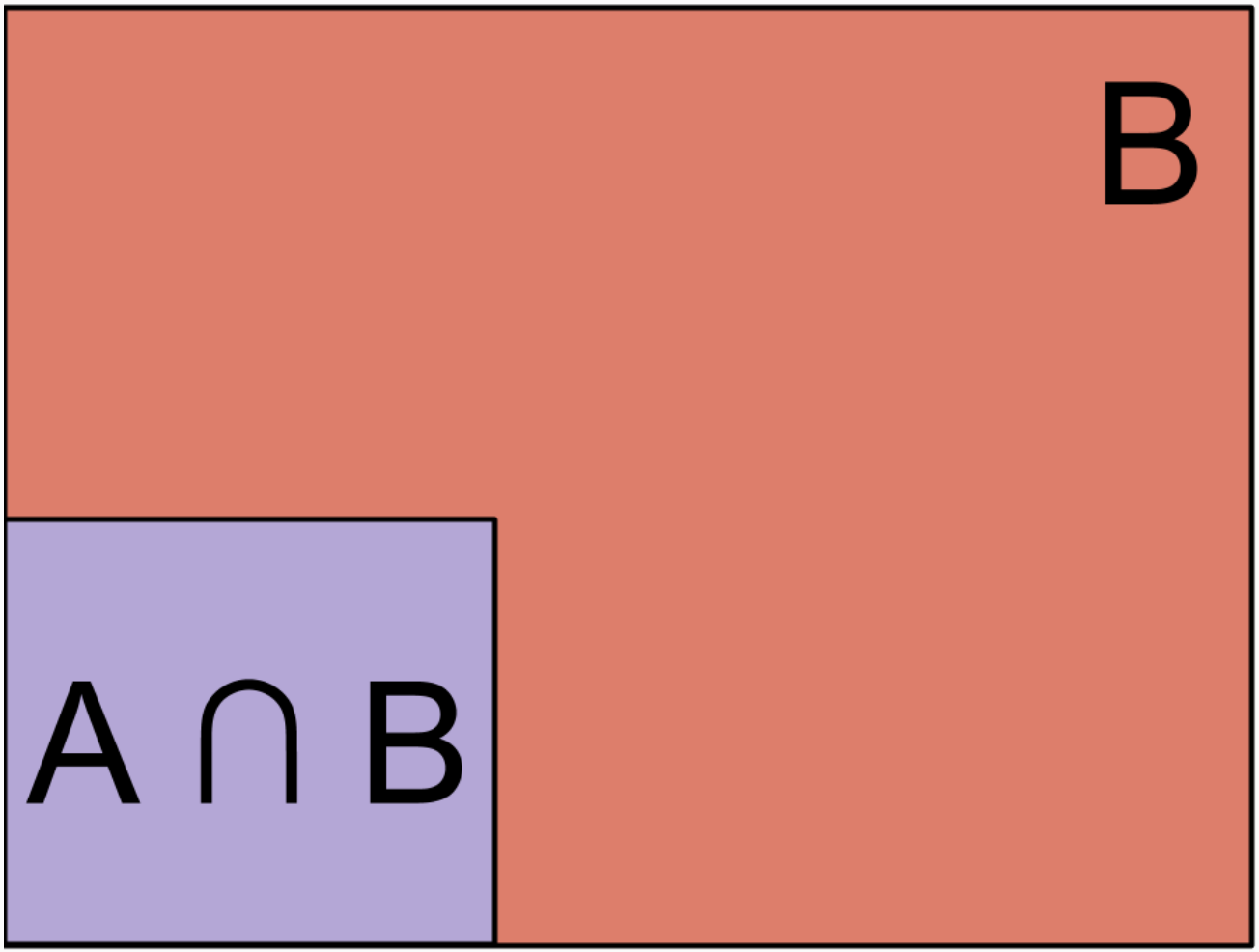
$$\mathbf{Pr}(Y = y|X = x)$$

Compute $\mathbf{Pr}(Y = 5|X = \mathit{female})$.
Compute $\mathbf{Pr}(Y = 5|X = \mathit{male})$.
Compute $\mathbf{Pr}(X = \mathit{female}|Y = 5)$.

conditional probability = rescaling



conditional probability = rescaling



conditional probability

Conditional probability is an axiom (definition) of probability theory. If $\mathbf{Pr}(X) > 0$

$$\mathbf{Pr}(Y|X) = \frac{\mathbf{Pr}(X, Y)}{\mathbf{Pr}(X)}$$

This implies that a joint probability can be factorized:

$$\mathbf{Pr}(X, Y) = \mathbf{Pr}(X)\mathbf{Pr}(Y|X)$$

and so on,

$$\mathbf{Pr}(X, Y, Z) = \mathbf{Pr}(X)\mathbf{Pr}(Y|X)\mathbf{Pr}(Z|X, Y).$$

relations between probabilities

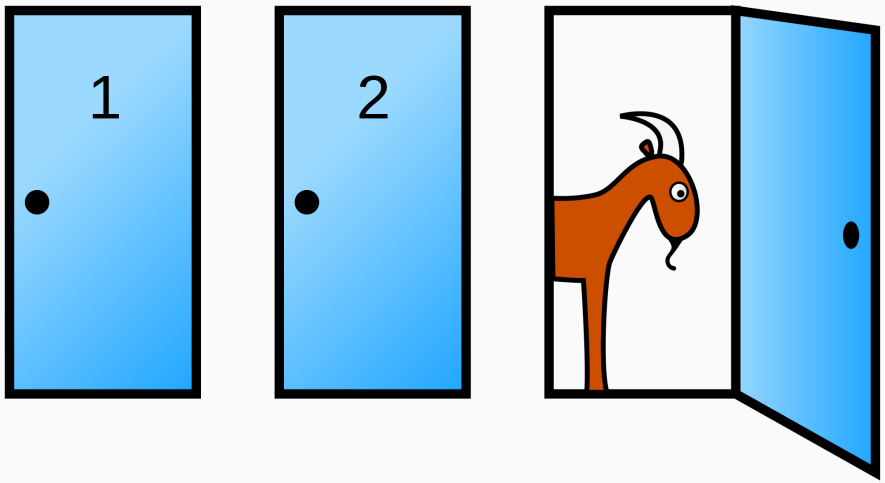
If A and B are disjoint outcomes (e.g. different sides of the same coin) then $\mathbf{Pr}(X = A, X = B) = ?$

If A is an event and B is the event that A did not occur then $\mathbf{Pr}(X = A) + \mathbf{Pr}(X = B) = ?$

For arbitrary events A and B which may intersect, how can we relate $\mathbf{Pr}(X = A) + \mathbf{Pr}(X = B)$ to $\mathbf{Pr}(X = A, X = B)$ in general?

If A and B are outcomes of independent events (e.g. results of flipping two different coins) then $\mathbf{Pr}(X_1 = A, X_2 = B) = ?$

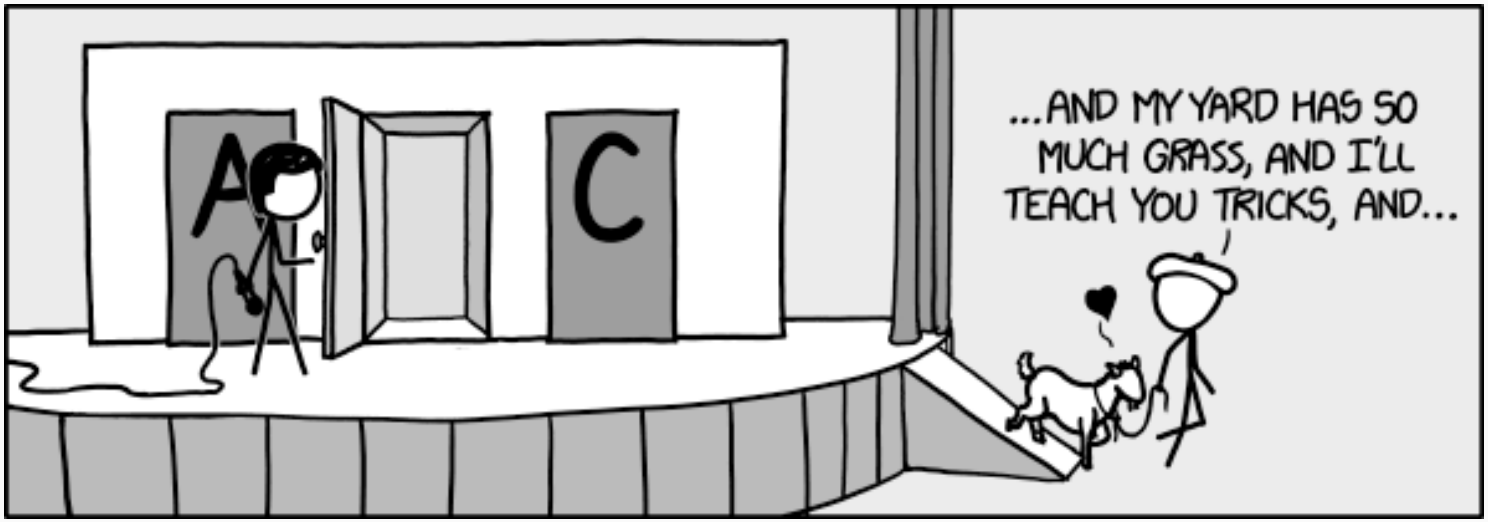
a puzzle: can you win a car?



1. There's a car behind one door and goats behind the others.
2. You choose one door and stand in front of it.
3. Monty opens a different door and reveals a goat.
4. Monty asks you if you want to choose the other door.

What do you answer?

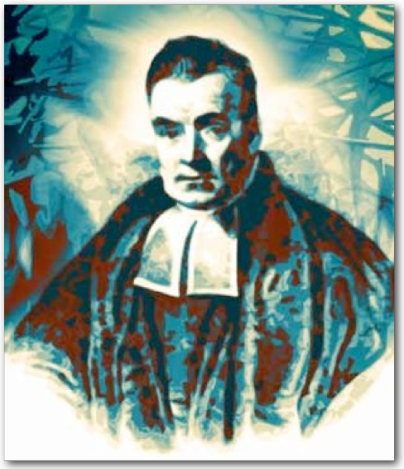
a puzzle

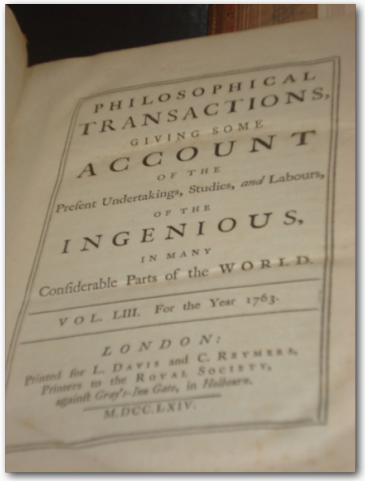


Why should you switch?

What's the probability you win a car if you switch?

rev. bayes





$$\Pr(X|Y) = \frac{\Pr(X)\Pr(Y|X)}{\Pr(Y)}$$

or

$$\Pr(X|Y) = \frac{\Pr(X,Y)}{\sum x \Pr(Y, X = x)}$$

Question 1

There is a school with 60% boys and 40% girls. The female students wear trousers or skirts in equal numbers; the boys all wear trousers. A passer-by sees a (random) student from a distance wearing trousers.

What is the probability that this student is a girl?

Question 2

You have 3 identical looking coins, two of them fair and the other a counterfeit that always lands heads. You pick one of them at random.

What is the prior probability that you chose the counterfeit coin?

You flip the coin three times and see that it lands heads each time.

What is the posterior probability that you chose the counterfeit coin?

practical exercise

Your friend has a bag of red and blue coins.
He draws a coin at random and then tosses the coin n times.
He repeats this procedure t times.

E.g. with $n = 3$ and $t = 5$ you might see
(B, H, H, H), (R, T, T, H), (R, H, T, T), (B, H, H, T), (R, H, T, T).

What parameters would you need to model this data?

How can you estimate them?

practical exercise

Run the python script coins_example.py

Fix TODOs.

How long do the estimates take to converge?

Does it matter how big each sample is (i.e. n)?

There is a solution in coins_example_solution.py

practical exercise

Your careless friend dropped the bag of coins in the bath.
The paint was not waterproof so now all the coins are white.
How would you estimate the parameters now?

i.e. you see only (H, H, H), (T, T, H), (H, T, T), (H, H, T), (H, T, T).

Look at the script hidden_coins_example.py.

Expectation-Maximization

maximum likelihood from observed data

We observe *i.i.d.* samples D of two random variables X and Z drawn from a distribution with parameters θ, i.e.

$$\Pr(D|\theta) = \prod_{(x,z)\in D} \Pr(X = x, Z = z|\theta)$$

To estimate the parameters we can maximize the likelihood of D or equivalently its logarithm.

$$\begin{aligned}
\hat{\theta}_{mle} &= \operatorname*{argmax}_{\theta} \Pr(D|\theta) \\
&= \operatorname*{argmax}_{\theta} \prod_{(x,z)\in D} \Pr(X = x, Z = z|\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{(x,z)\in D} \log \Pr(X = x, Z = z|\theta)
\end{aligned}$$

maximum likelihood: coins

We observed a sample D drawn from $(x, y) \in (X, Z)$ where $X \in \{H, T\}$, $Z = \{A, B\}$. Each observation was labeled so,

$$
\begin{aligned}
\hat{\theta}_{mle} &= \operatorname*{argmax}_{\theta} \mathbf{Pr}(D|\theta) \\
&= \operatorname*{argmax}_{\theta} \prod_{(x,z)\in D} \mathbf{Pr}(X = x, Z = z|\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{(x,z)\in D} \log \mathbf{Pr}(X = x, Z = z|\theta) \\
&= \operatorname*{argmax}_{\theta} \sum_{(x,z)\in (X,Z)} \#(X = x, Z = z) \log \mathbf{Pr}(X = x, Z = z|\theta)
\end{aligned}
$$

from mle to em

Let's reformulate the expression for *mle* estimation.

$$\begin{aligned}\hat{\theta}_{mle} &= \operatorname*{argmax}_{\theta} \sum_{(x,z)\in D} \log \mathbf{Pr}(X = x, Z = z|\theta) \\ &= \operatorname*{argmax}_{\theta} \sum_{(x,z)\in D} \sum_{y\in\{A,B\}} \delta(z, y) \log \mathbf{Pr}(X = x, Z = z|\theta)\end{aligned}$$

where $\delta(x, y) = 1 \iff x = y$ otherwise 0.

hidden data parameter estimation

We observed a sample D drawn from $(x, z) \in (X, Z)$ where $X \in \{H, T\}$, $Z = \{A, B\}$. This time Z is hidden.

$$\hat{\theta}_{mle} = \operatorname*{argmax}_{\theta} \sum_{(x,z) \in D} \sum_{y \in \{A,B\}} \delta(z, y) \log \mathbf{Pr}(X = x, Z = z | \theta)$$

Replace $\delta(z, y) \in \{0, 1\}$ by our best guess $\mathbf{Pr}(Z_t = z | X = x, \theta_t)$.

$$\hat{\theta}_{t+1} = \operatorname*{argmax}_{\theta} \sum_{x \in D} \sum_{z \in \{A,B\}} \mathbf{Pr}(Z_t = z | X = x, \theta_t) \log \mathbf{Pr}(X = x, Z_t = z | \theta_t)$$

This term is known as the *expected log-likelihood*.

expectation maximization

- Initialize the parameters θ_0 somehow (randomly?)
- E-step: Compute $\mathbf{Pr}(Z_t|X, \theta_t)$ i.e. our best guess of the hidden data Z given our current parameters. (Think of $\mathbf{Pr}(Z_t|X, \theta_t)$ as a fractional count of Z at time t.)
- M-step: Update the parameters θ_{t+1} to maximize the expected log-likelihood.
- Iterate until the expected log-likelihood stops increasing.

Intuition: if we knew θ we could just infer Z (usually), likewise if we knew Z we could just estimate θ (you did this). Since we don't know either, just guess and iteratively improve.

em maximizes a bound on the observed log-likelihood

$$\underbrace{\log \Pr(X|\theta)}_{\text{observed}} = \log \sum_Z \Pr(X, Z|\theta) \tag{1}$$

$$= \log \sum_Z q(Z) \frac{\Pr(X, Z|\theta)}{q(Z)} \tag{2}$$

$$\geq \sum_Z q(Z) \log \frac{\Pr(X, Z|\theta)}{q(Z)} \tag{3}$$

$$= \underbrace{\sum_Z q(Z) \log \Pr(X, Z|\theta)}_{\text{expected}} - \underbrace{\sum_Z q(Z) \log q(Z)}_{\text{entropy}} \tag{4}$$

$q(Z)$ doesn't depend on θ so we can ignore it in the M-step.

em maximizes a bound on the observed log-likelihood

$$\underbrace{\log \mathbf{Pr}(X|\theta)}_{\text{observed}} = \log \sum_Z \mathbf{Pr}(X, Z|\theta) \tag{5}$$

$$\geq \sum_Z q(Z) \log \frac{\mathbf{Pr}(X|\theta)(Z|X, \theta)}{q(Z)} \tag{6}$$

$$= \sum_Z q(Z) \log \mathbf{Pr}(X|\theta) - \sum_Z q(Z) \log \frac{q(Z)}{\mathbf{Pr}(Z|X, \theta)} \tag{7}$$

$$= \log \mathbf{Pr}(X|\theta) - KL(q(Z)||\mathbf{Pr}(Z|X, \theta)) \tag{8}$$

Choosing $q(Z) = \mathbf{Pr}(Z|X, \theta)$ we make bound tight in the E-step.

considerations for using em

- Efficient inference: to repeatedly re-estimate the posterior distribution over hidden variables in the E-step.
- Efficient parameter updates: to repeatedly update parameters in the M-step.

What do those constraints mean for our models?

hard em variants

- (K-means): Assign each data point X to most probable class.

$$Z_t = \underset{z \in \mathcal{Z}}{\operatorname{argmax}}\, Pr(Z_t = z | X, \theta_t)$$

- (Gibbs sampling): Sample a label for each data point X from the posterior distribution.

$$Z_t \sim_{z \in \mathcal{Z}} \mathbf{Pr}(Z_t = z | X, \theta_t)$$

What might be the advantages of these variants?