Bayes' Rule for Random Variables (Also Called Bayes' Theorem for Random Variables

6.008.1x - Bayes' Rule for Random Variables (Also Called Bayes' Theorem for Rando

0:00 / 0:00 1.25x

Video
Download video file

Transcripts
Download SubRip (.srt) file
Download Text (.txt) file

These notes cover roughly the same content as the video:

BAYES' THEOREM FOR RANDOM VARIABLES (COURSE NOTES)

In inference, what we want to reason about is some unknown random variable X, where we get to observe some other random variable Y, and we have some model for how X and Y relate. Specifically, suppose that we have some “prior" distribution p_X for X; this prior distribution encodes what we believe to be likely or unlikely values that X takes on, *before we actually have any observations*. We also suppose we have a “likelihood" distribution $p_{Y|X}$.

After observing that Y takes on a specific value y, our “belief" of what X given $Y = y$ is now given by what's called the “posterior" distribution $p_{X|Y}(\cdot \mid y)$. Put another way, we keep track of a probability distribution that tells us how plausible we think different values X can take on are. When we observe data Y that can help us reason about X, we proceed to either upweight or downweight how plausible we think different values X can take on are, making sure that we end up with a probability distribution giving us our updated belief of what X can be.

Thus, once we have observed $Y = y$, our belief of what X is changes from the prior p_X to the posterior $p_{X|Y}(\cdot \mid y)$.

Bayes' theorem (also called Bayes' rule or Bayes' law) for random variables explicitly tells us how to compute the posterior distribution $p_{X|Y}(\cdot \mid y)$, i.e., how to weight each possible value that random variable X can take on, once we've observed $Y = y$. Bayes' theorem is the main workhorse of numerous inference algorithms and will show up many times throughout the course.

Bayes' theorem: Suppose that y is a value that random variable Y can take on, and $p_Y(y) > 0$. Then

$$p_{X|Y}(x \mid y) = \frac{p_X(x)p_{Y|X}(y \mid x)}{\sum_{x'} p_X(x')p_{Y|X}(y \mid x')}$$

for all values x that random variable X can take on.

Important: Remember that $p_{Y|X}(\cdot \mid x)$ could be undefined but this isn't an issue since this happens precisely when $p_X(x) = 0$, and we know that $p_{X,Y}(x, y) = 0$ (for every y) whenever $p_X(x) = 0$.

Proof: We have

$$p_{X|Y}(x \mid y) \overset{(a)}{=} \frac{p_{X,Y}(x,y)}{p_Y(y)} \overset{(b)}{=} \frac{p_X(x)p_{Y|X}(y \mid x)}{p_Y(y)} \overset{(c)}{=} \frac{p_X(x)p_{Y|X}(y \mid x)}{\sum_{x'} p_{X,Y}(x', y)} \overset{(d)}{=} \frac{p_X(x)p_{Y|X}(y \mid x)}{\sum_{x'} p_X(x')p_{Y|X}(y \mid x')}$$

where step (a) uses the definition of conditional probability (this step requires $p_Y(y) > 0$), step (b) uses the product rule (recall that for notational convenience we're not separately writing out the case when $p_X(x) = 0$), step (c) uses the formula for marginalization, and step (d) uses the product rule (again, for notational convenience, we're not separately writing out the case when $p_X(x') = 0$). $\square$