E6892 Bayesian Models for Machine Learning

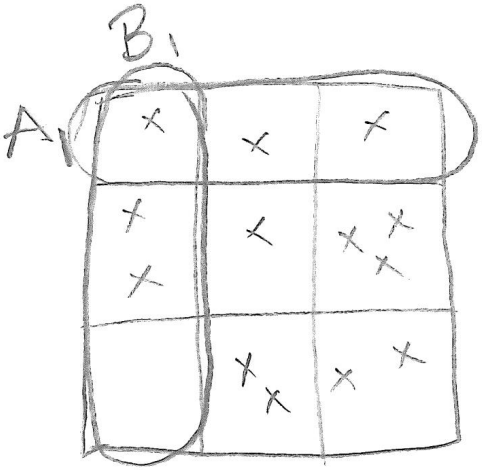
Columbia University, Fall 2015

Lecture 1, 9/10/2015

Instructor: John Paisley

- "Bayes rule" pops out of basic manipulations of probability distributions. Let's reach it through a very simple example.

Example



Call this entire space Ω

A_i is the ith column (defined arbitrarily)

B_i is the ith row (also defined arbitrarily)

- We have points lying in this space. We pick one of these points uniformly at random.
- In this case, calculating probabilities is simply a matter of counting.

$$P(x \in A_1) = \frac{\#A_1}{\#\Omega}, \qquad P(x \in B_1) = \frac{\#B_1}{\#\Omega}$$

- What about $P(x \in A_1 | x \in B_1)$? This is the probability that $x \in A_1$ given that I know $x \in B_1$. This is called a *conditional probability*.

$$P(x \in A_1 | x \in B_1) = \frac{\#(A_1 \cap B_1)}{\#B_1} = \frac{\#(A_1 \cap B_1)}{\#\Omega}\frac{\#\Omega}{\#B_1} = \frac{P(x \in A_1 \,\&\, x \in B_1)}{P(x \in B_1)}$$

- We've simply multiplied and divided by the same thing, but already we've made a general statement.

A more general statement

- Let A and B be two events, then

$$P(A|B) = \frac{P(A,B)}{P(B)} \qquad \Rightarrow \qquad P(A|B)P(B) = P(A,B)$$

- We have some names for these terms,

$$\begin{aligned} P(A|B) &: \text{ conditional probability distribution} \\ P(A,B) &: \text{ joint probability distribution} \\ P(B) &: \text{ marginal probability distribution} \end{aligned}$$

- This last one could be tricky, since it's also just "the probability of B." However, we can use the same counting approach as before to interpret $P(B)$ as a distribution arrived at by integrating (or marginalizing) something out. From the previous example,

$$P(B) = \frac{\#B}{\#\Omega} = \frac{\sum_{i=1}^{3} \#(A_i \cap B)}{\#\Omega} = \sum_{i=1}^{3} \frac{\#(A_i \cap B)}{\#\Omega} = \sum_{i=1}^{3} P(A_i, B)$$

- Side note: In general, the summation and each A_i have a very strict requirement. Each A_i has to be disjoint (no overlaps) and the union of all the A_i has to equal Ω (the entire space).

Getting to Bayes rule

- We're a few easy steps away. We showed that

$$P(A,B) = P(A|B)P(B)$$

By symmetry we could just as easily have shown that

$$P(A,B) = P(B|A)P(A)$$

And therefore,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_i P(A_i, B)} = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

- This equality is called *Bayes rule*. As you can see, it has a few ways it can be written.

Bayes rule

- And so we have that $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

- These values each have a name:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- Imagine that we don't know A, but we get some information about it in the form of B. Bayes rule tells us a principled way to incorporate this information in our belief about A.

Example (medical test)

- We have two binary values, A and B:

$$A = \begin{cases} 1 & \text{person has disease} \\ 0 & \text{no disease} \end{cases} \qquad B = \begin{cases} 1 & \text{test for disease "positive"} \\ 0 & \text{test is "negative"} \end{cases}$$

- A person tests "positive" for the disease. What's the probability he has it?

- We want $P(A = 1|B = 1)$. Does Bayes rule help? Bayes rule says that

$$\begin{aligned} P(A = 1|B = 1) &= \frac{P(B = 1|A = 1)P(A = 1)}{P(B = 1)} \\ &= \frac{P(B = 1|A = 1)P(A = 1)}{P(B = 1|A = 1)P(A = 1) + P(B = 1|A = 0)P(A = 0)} \end{aligned}$$

- Image we can estimate that

$$P(B = 1|A = 1) = 0.95, \quad P(B = 1|A = 0) = 0.05, \quad P(A = 1) = 0.01 = 1 - P(A = 0)$$

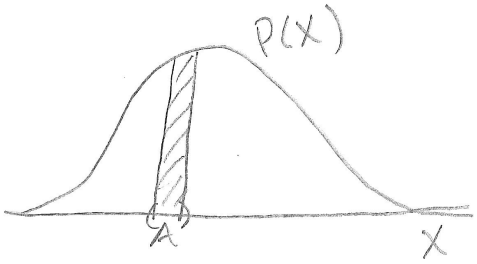
Then plugging in, we have that $P(A = 1|B = 1) = 0.16$

Continuous space

- We've been talking about discrete distributions so far. That is, the number of possible values the unknowns can take is finite.

- When values are in a continuous space, we switch to continuous distributions.

Example



$x \in \mathbb{R}$, $p(x)$ is its density (represented by the switch to lower case)

$p(x) \geq 0$ and $\int p(x)dx = 1$

$P(x \in A) = \int_A p(x)dx$

- The same rules apply as for discrete random variables

- $p(x|y) = \dfrac{p(x,y)}{p(y)}, \qquad p(y) = \displaystyle\int p(x,y)dx$
- $p(x|y)p(y) = p(x,y) = p(y|x)p(x)$

- This leads to *Bayes rule* for continuous random variables

$$p(y|x) \;=\; \frac{p(x|y)p(y)}{p(x)} \;=\; \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

- The difference is that we are dealing with continuous functions.

Bayesian modeling

- Applying Bayes rule to the unknown variables of a data modeling problem is called Bayesian modeling.
- In a simple, generic form we can write this process as

  $x \sim p(x|y) \;\leftarrow$ The data-generating distribution. This is the model of the data.

  $y \sim p(y) \;\leftarrow$ The model prior distribution. This is what we think about y a priori.

- We want to learn y. We write that

$$p(y|x) \propto p(x|y)p(y)$$

- In the above line, we don't know $p(y|x)$ and want to calculate it. It answers the question "Having seen x, what can we say about y?" The right two terms $p(x|y)p(y)$ we do know because we have *defined* it. The symbol $\propto$ is read as "is distributed as."
- This is the general form of the problems discussed in this class. However, it is non-trivial because
  1. $p(x|y)$ can be quite complicated
  2. $p(y|x)$ can be intractable because the integral in the normalizing constant doesn't have a closed-form solution, thus requiring an algorithm to approximate

- These two issues will make up the focus of this class: Defining various models on the structure of the data-generating phenomenon, and defining inference algorithms for learning the posterior distribution of that model's variables.

Simple example: beta-Bernoulli model

- We have a sequence of observations $X_1, \ldots, X_N$, where $X_i = 1$ indicates "success" and $X_i = 0$ indicates "failure." Think of them as results of flipping a coin.
- We hypothesize that each X_i is generated by flipping a biased coin, where $P(X_i = 1|\pi) = \pi$.

- We further assume the X_i are *independent and identically distributed* (iid). This means the X_i are *conditionally independent* given π. Mathematically this means we can write

$$P(X_1, \ldots, X_N|\pi) = \prod_{i=1}^{N} P(X_i|\pi)$$

- Since $P(X_i|\pi) = \pi^{X_i}(1-\pi)^{1-X_i}$, we can write

$$P(X_1, \ldots, X_N|\pi) = \prod_{i=1}^{N} \pi^{X_i}(1-\pi)^{1-X_i}$$

- We are interested in the posterior distribution of π given $X_1, \ldots, X_N$, i.e.,

$$\begin{aligned} P(\pi|X_1, \ldots, X_N) &\propto P(X_1, \ldots, X_N|\pi)p(\pi) \\ &\propto \prod_{i=1}^{N} \pi^{X_i}(1-\pi)^{1-X_i}p(\pi) \end{aligned}$$

- We've come across the next significant problem: What do we set $p(\pi)$ to?

A first try

- Let $p(\pi) = Uniform(0,1) \;\Rightarrow\; p(\pi) = \mathbb{1}(0 \leq \pi \leq 1)$
- Then by Bayes rule,

$$p(\pi|X_1, \ldots, X_N) = \frac{\pi^{(\sum_i X_i+1)-1}(1-\pi)^{(N-\sum_i X_i+1)-1}}{\int_0^1 \pi^{\sum_i X_i}(1-\pi)^{N-\sum_i X_i} d\pi}$$

- The normalizing constant is tricky, but fortunately mathematicians have solved it,

$$p(\pi|X_1, \ldots, X_N) = \frac{\Gamma(N+2)}{\Gamma(1+\sum_i X_i)\Gamma(1+N-\sum_i X_i)} \pi^{\sum_i X_i+1-1}(1-\pi)^{N-\sum_i X_i+1-1}$$

  ($\Gamma(\cdot)$ is called the "gamma function")

- This is a very common distribution called a *beta distribution*:

$$Beta(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1}$$

- In the case of the above posterior distribution, $a = 1 + \sum_i X_i$ and $b = 1 + N - \sum_i X_i$.
- Notice that when $a = b = 1$, $Beta(1,1) = Uniform(0,1)$ which was the prior we chose.

A conjugate prior

- The beta distribution looks a lot like the likelihood term. Also, because it has the $Uniform(0,1)$ distribution as a special case, it would give us a wider range of prior beliefs that we could express. What if we try the beta distribution as the prior for π?

$$\begin{aligned} p(\pi|X_1,\ldots,X_N) &\propto p(X_1,\ldots,X_N|\pi)p(\pi) \\ &\propto \left[\pi^{\sum_i X_i}(1-\pi)^{N-\sum_i X_i}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}\right] \\ &\propto \pi^{a+\sum_i X_i-1}(1-\pi)^{b+N-\sum_i X_i-1} \end{aligned}$$

- We can notice a few things here:

  1. I threw away $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ in the last line. This is because it doesn't depend on π, and so it can be absorbed in the normalizing constant. In other words, the normalizing constant is the integral of the numerator. If we divide the second line by this integral, we can immediately cancel out $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$. The result is the third line divided by the integral of the third line. The only reason to do something like this is for convenience and to make things look simpler.

  2. In this case, we know what to divide this by to make it a probability distribution on π. The last line is proportional to a $Beta(a+\sum_i X_i, b+N-\sum_i X_i)$ distribution.

- Therefore, $p(\pi|X_1,\ldots,X_N) = Beta(a+\sum_i X_i, b+N-\sum_i X_i)$.

- Notice that for this problem, when we select the prior $p(\pi)$ to be a beta distribution, we find that the posterior distribution is also a beta distribution with different parameters. This is because the beta distribution is a *conjugate prior* for this problem.

- We say that a distribution is a "conjugate prior" for a particular variable in a likelihood, or that it is "conjugate to the likelihood," if the posterior is in the same distribution family as the prior, but has updated parameters. Conjugate priors are very convenient, since we only need to collect *sufficient statistics* from the data in order to transition from the prior to the posterior. For example, in the problem above we only need to count the total number of "successes" ($\sum_i X_i$) and "failures" ($N-\sum_i X_i$).

- will see many conjugate priors in this course and how they can result in fast inference algorithms for learning models.

What do we gain by being Bayesian?

- For the previous modeling problem with a beta prior, consider the expectation and variance of π under the posterior distribution.

$$\mathbb{E}[\pi] = \frac{a+\sum_i X_i}{a+b+N}, \qquad Var(\pi) = \frac{(a+\sum_i X_i)(b+N-\sum_i X_i)}{(a+b+N)^2(a+b+N-1)}$$

- Notice that as N increases,

    1. The expectation converges to the empirical success rate.

    2. The variance decreases like $1/N$, i.e., it's going to zero.

- Compare this with *maximum likelihood*, in which we seek a point estimate (on specific value) of π to maximize the likelihood term only

$$\pi = \arg\max_{\pi} \; p(X_1, \ldots, X_N|\pi) \; = \; \frac{1}{N} \sum_{i=1}^{N} X_i$$

- The Bayesian approach is capturing our uncertainty about the quantity we are interested in. Maximum likelihood does not do this.

- As we get more and more data, the Bayesian and ML approaches agree more and more. However, Bayesian methods allow for a smooth transition from uncertainty to certainty.