CS 281: Advanced Machine Learning

taught by Sasha Rush

Fall 2017

Contents

CS281: Advanced ML **September 6, Fall 2017**

Lecture 1: Discrete Models

Lecturer: Sasha Rush *Scribes: Anna Sophie Hilgard*

- Take values from a countable set, e.g. {0,1}, {cold, flu, asthma}
- We will use simple discrete models to develop our tactics such as marginalization and conditioning. Today: coins.

1.1 Bernoulli model

The likelihood is of the form $p(\text{heads}) = \theta$.

Easy Prior

Assume we know the coin came from one of 3 unknown manufacturers (Later, we'll have mixture model estimation, but for now assume these probabilities come from an oracle.)
$\theta = 0.4$ with probability .1
$\theta = 0.5$ with probability .8
$\theta = 0.6$ with probability .1
$p(\theta) = 0.1 \cdot \delta(\theta = 0.4) + 0.8 \cdot \delta(\theta = 0.5) + 0.1 \cdot \delta(\theta = 0.6)$

Likelihood

Likelihood: $p(\text{data}|\text{parameters})$
Ex: $p(\text{coin\_flips}|\theta) = \text{Bin}(N_1|N, \theta) = \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N-N_1}$
Where $N = N_0 + N_1 =$ number of flips
Note that the last term, the "score", is the only term that depends on θ.

1.2 Inference

Inference 1: $p(\theta|x)$ $(x = N_0, N_1)$

Maximum Likelihood Estimation (MLE)

$\theta_{MLE} = \text{argmax}_\theta\ p(N_0, N_1|\theta) = \text{argmax}_\theta \log\left[p(N_0, N_1|\theta)\right]$
$\theta_{MLE} = \text{argmax}_\theta \log\binom{N}{N_1} + N_1 \log\theta + N_0 \log(1-\theta)$
Because the first term is not a function of θ, we can ignore it.
$\frac{d}{d\theta} = \frac{N_1}{\theta} + \frac{N_0}{1-\theta} \cdot (-1) \rightarrow \theta_{MLE} = \frac{N_1}{N_0+N_1}$

- Note that Inference != Decision Making. If we asked you to make a bet on the coin, based on this you could either

  1. Always take heads if $\theta > .5$. In this case, $p(win) = \theta$
  2. Take heads with probability $= \theta$. In this case, $p(win) = \theta^2 + (1-\theta)^2$ [p(is heads) * p(choose heads) + ...]

  For $\theta = .6$, the win probabilities are .6 and .52, respectively.

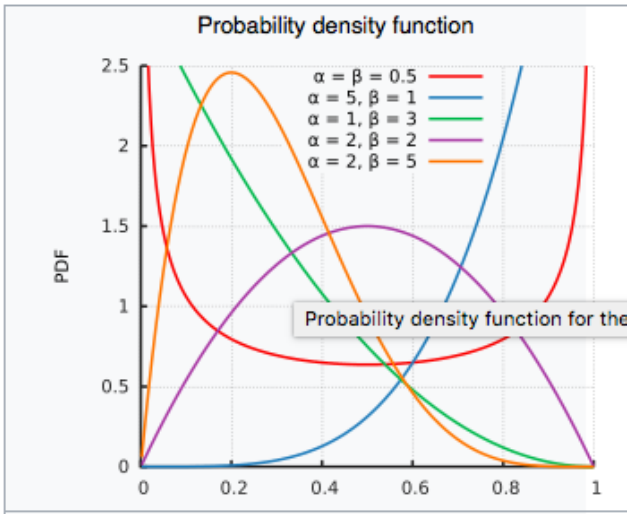

Figure 1.1: Beta Params

Maximizing the Posterior (MAP)

Bayes Rule : $p(\theta|data) \propto p(data|\theta)p(\theta)$

- Posterior: $p(\theta|x)$
- Likelihood: $p(x|\theta)$
- Prior: $p(\theta)$

$\theta_{MAP} = \text{argmax}_\theta\, p(\theta|x) = \text{argmax}_\theta \log\left[p(x|\theta)p(\theta)\right]$

Ex:
$p(\theta = 0.4|N_0, N_1) \propto \binom{N}{N_1}(.4)^{N_1}(1-.4)^{N_0}(0.1)$
$p(\theta = 0.45|N_0, N_1) = 0$ (Due to the sparsity of the prior)

MAP = MLE when we have a uniform prior

Full Posterior

- Partition/Marginal: $p(N_0, N_1) = \int_\theta p(N_0, N_1, \theta)$. Note this is evil and hard to compute

$p(\theta|N_0, N_1) = \frac{p(x|\theta)p(\theta)}{p(N_0,N_1)}$

Beta Prior

Beta Distribution: $p(\theta|\alpha_0, \alpha_1) = \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}$

From the image of the beta function for different parameters, we can see that it can ether be balanced, humped to one side, or tend toward infinity on one side or the other.

With a beta prior: $p(\theta|N_0, N_1) = \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1}\theta^{N_1}(1-\theta)^{N_0}\cdot$(constant normalizer w.r.tθ)

The key insight is that we get additive terms in the exponent and the resulting distribution looks like another beta. The prior "counts" (pseudocounts) from the hyperparameters can be interpreted as counts we have beforehand

$p(\theta|N_0, N_1) = \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1} \cdot (\text{constant})$
To make this distribution sum to 1, use the known beta normalizer
$p(\theta|N_0, N_1) = \frac{\Gamma(\alpha_0+\alpha_1+N_0+N_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(N_0)\Gamma(N_1)}\theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1}$
$p(\theta|N_0, N_1) \sim \text{Beta}(\theta|N_0+\alpha_0, N_1+\alpha_1)$ is the posterior distribution.

- The mode of the Beta gives us back the MAP, but we additionally now have information about the shape.
- What does the prior that tends to infinity at 1 imply? That in the absence of other information, the coin is definitely heads.

Predictive Distribution

$p(\hat{x}|N_0, N_1) = \int_\theta p(x|\theta, N_0, N_1)p(\theta|N_0, N_1)d\theta$
$p(\hat{x}|N_0, N_1) = \int_\theta \theta p(\theta|N_0, N_1)d\theta$
$p(\hat{x}|N_0, N_1) = \mathbb{E}_{\theta\sim p(\theta|N_0,N_1)}\theta$ (The expectation under the posterior of θ. This is the mean of the Beta distribution. Feel free to prove it as an exercise!)

Normalizing Term / Marginal (Likelihood)

$p(N_0, N_1) = \int_\theta p(x_1, \ldots x_n|\theta)p(\theta)d\theta$
$p(N_0, N_1) = \int_\theta \frac{\Gamma(\alpha_1+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1+N_1-1}(1-\theta)^{\alpha_0+N_0-1}$
The first term can be moved outside, as it does not depend on θ. Then we know what the integral of the rest should be because we know how to make the distribution sum to 1.
$p(N_0, N_1) = \frac{\Gamma(\alpha_1+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\frac{\Gamma(N_0+\alpha_0)\Gamma(N_1+\alpha_1}{\Gamma(N_0+N_1+\alpha_0+\alpha_1}$

1.3 Extensions on the Coin Flip Model: Super Coins

- Many coins and they're correlated (Models of binary data)
- Many-sided coins aka dice (Models of categorical data)

Bernoulli

$\text{Ber}(x|\theta) = \theta^x(1-\theta)^x$

Categorical

$\text{Cat}(x|\theta) = \prod_k \theta_k^{x_k}$

Multinomial

$\text{Multi}(x|\theta) = \frac{(\sum x_k)!}{\prod_k x_k!}\prod_k \theta_k^{x_k}$

Dirichlet

$\text{Dir}(x|\alpha) = \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)}\prod_k \theta_k^{\alpha_k-1}$

1.4 Example notebook

See Beta.ipynb