# CS 281: Advanced Machine Learning

taught by Sasha Rush

Fall 2017

## Contents

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

# Lecture 1: Discrete Models

*Lecturer: Sasha Rush* *Scribes: Anna Sophie Hilgard, Diondra Peck*

- Discrete models take values from a countable set, e.g. {0,1}, {cold, flu, asthma} and are simpler than continuous models.
- We will use simple discrete models to develop our tactics such as marginalization and conditioning.
- Today, we will focus coins as a real-world example.

## 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).

1. $\theta = 0.4$ with probability .1
2. $\theta = 0.5$ with probability .8
3. $\theta = 0.6$ 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 = $p(\text{data}|\text{parameters})$. For the coin example,

$$p(\text{coin flips}|\theta) = \text{Bin}(N_1|N, \theta) = \binom{N}{N_1}\theta^{N_1}(1-\theta)^{N-N_1} \quad \text{where } N = N_0 + N_1 = \text{number of flips}$$

Note that the last two terms, the "score", is our focus since they are the only terms that depend on $\theta$. The first term normalizes the distribution.

## 1.2 Inference

Inference 1: $p(\theta|x)$ ($x \in N_0, N_1$). How can we estimate $\theta$?

#### 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 $\theta$, 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 $\neq$ 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, $p(\text{win}) = \theta$
2. Take heads with probability = $\theta$. In this case, $p(\text{win}) = \theta^2 + (1-\theta)^2$ [p(is heads) * p(choose heads) + ...]

If $\theta$ = 0.6, for option 1, p(win) = $\theta$ = 0.6. For option 2, p(win) = $\theta^2 + (1-\theta)$ = 0.52. In this case, the additional information used in the calculation of option 2 does not result in a better decision.

**Maximizing the Posterior (MAP)**

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

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

$$\theta_{MAP} = \text{argmax}_\theta \; p(\theta|\text{x}) = \text{argmax}_\theta \log\left[p(\text{x}|\theta)p(\theta)\right] \quad \text{from Bayes' Rule: } p(\theta|\text{x}) \propto p(\text{x}|\theta)p(\theta)$$

Consider an example:

$$p(\theta = 0.4|N_0, N_1) \propto \binom{N}{N_1}(.4)^{N_1}(1-.4)^{N_0}(0.1)$$

$$p(\theta = 0.45|N_0, N_1) = 0 \quad \text{Due to the sparsity of the prior - similar result for } \theta = 0.5 \text{ and } 0.6$$

$\theta_{MAP} = \theta_{MLE}$ when we have a uniform prior since the MLE calculation does not explicitly factor a prior into its calculation.

**Full Posterior**

Partition or Marginal Likelihood: $p(N_0, N_1) = \int_\theta p(N_0, N_1, \theta)$.

$$p(\theta|N_0, N_1) = \frac{p(x|\theta)p(\theta)}{p(N_0, N_1)} \qquad \text{Note that } p(N_0, N_1) \text{ is a very difficult term to compute.}$$

**Beta Prior**

$$p(\theta|\alpha_0, \alpha_1) = \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \qquad \text{support} \in [0, 1]$$

From the image of the beta function for different parameters, we can see that it can ether be balanced, skewed to one side, or tend toward infinity on one side.
With a beta prior:

$$p(\theta|N_0, N_1) = \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1}\theta^{N_1}(1-\theta)^{N_0} \cdot (\text{constant normalization term w.r.t } \theta)$$

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.

$$p(\theta|N_0, N_1) = \frac{\Gamma(\alpha_0 + \alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{N_1 + \alpha_1 - 1}(1-\theta)^{N_0 + \alpha_0 - 1} \cdot (\text{constant normalization term w.r.t } \theta)$$

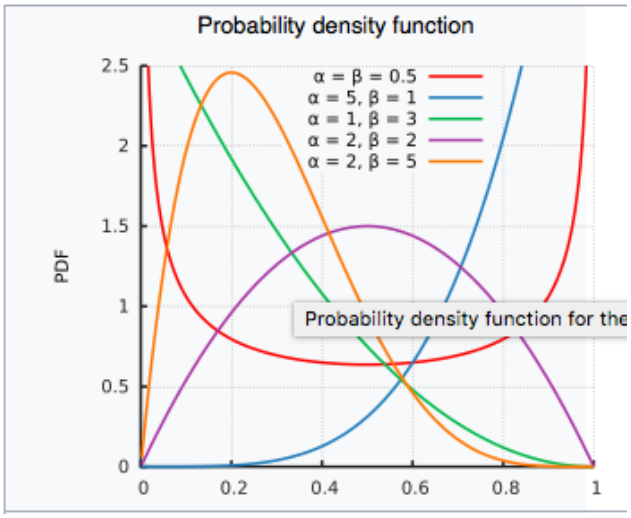


*Figure 1.1: Beta Params*

To make this distribution sum to 1, use the known beta normalizer

$$p(\theta|N_0, N_1) = \frac{\Gamma(\alpha_0 + \alpha_1 + N_0 + N_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(N_0)\Gamma(N_1)}\theta^{N_1+\alpha_1-1}(1-\theta)^{N_0+\alpha_0-1} \sim \text{Beta}(\theta|N_0+\alpha_0, N_1+\alpha_1) \quad \text{(posterior)}$$

The mode of the Beta gives us back $\theta_{MAP}$, but with additional information about the shape of the distribution. 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**

$$\begin{aligned} p(\hat{x}|N_0, N_1) &= \int_\theta p(x|\theta, N_0, N_1)p(\theta|N_0, N_1)d\theta \\ &= \int_\theta \theta p(\theta|N_0, N_1)d\theta \\ &= \mathbb{E}_{\theta\sim p(\theta|N_0,N_1)}\theta \end{aligned}$$

This is the expectation under the posterior of $\theta$ which is the mean of the Beta distribution. Feel free to prove this as an exercise.

**Marginal Likelihood**

$$\begin{aligned} p(N_0, N_1) &= \int_\theta p(x_1, \dots x_n|\theta)p(\theta)d\theta \\ &= \int_\theta \frac{\Gamma(\alpha_1+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\theta^{\alpha_1+N_1-1}(1-\theta)^{\alpha_0+N_0-1} \end{aligned}$$

The first term can be moved outside, as it does not depend on $\theta$. After introducing our normalization term and making the distribution insidesum 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 correlated coins: models of binary data, important for discrete graphical models
- Many-sided coins aka dice: models of categorical data, generalization of Bernoulli

## 1.4 Other Distributions

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

$$\text{Categorical}(x|\theta) = \prod_k \theta_k^{x_k} \qquad \text{generalization of Bernoulli}$$

$$\text{Multinomial}(x|\theta) = \frac{(\sum x_k)!}{\prod_k x_k!} \prod_k \theta_k^{x_k} \qquad \text{generalization of Binomial}$$

$$\text{Dirichlet}(x|\alpha) = \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \qquad \text{generalization of Beta, often used as a prior}$$

Note that the Dirichlet distribution is the conjugate prior of the Categorical and Multinomial distributions.

## 1.5 Example notebook

See Beta.ipynb