Probabilistic Machine Learning:
4. Beta-binomial model

Tomasz Kajdanowicz, Przemysław Kazienko (substitution)

Department of Computational Intelligence
Wroclaw University of Science and Technology

The presentation was inspired by Chapter 3 of Kevin Murphy book "Machine Learning A Probabilistic Perspective", 2012, MIT.
Apropriate agreements to propagate his ideas has been acquired.

Beta-binomial model

"Number game"

- inferring a distribution over a discrete variable drawn from a finite hypothesis space
- given a series of discrete observations
- computations particularly simple: sum, multiplication and division

What if, like in many applications, the unknown parameters are continuous?

- the hypothesis space is subset of $\mathbb{R}^K$, where K is the number of parameters
- replace sums with integrals

Coin toss example

The problem:

- inferring the probability that a coin shows up heads
- given a series of observed coin tosses

Might seem trivial, but

- this model forms the basis of many of the methods
- historically important, since it was the example which was analyzed in Bayes' original paper of 1763

Recipe of specifying the model

Define

- likelihood
- prior

and derive

- posterior
- posterior predictive

The problem

Let's consider a single binary random variable:

- $X_i \sim Bern(\theta)$
- $X_i = 1$ represents "heads", $X_i = 0$ represents "tails"
- $\theta \in [0, 1]$ is the parameter (probability of heads)
- $p(X_i = 1|\theta) = \theta, \;\; p(X_i = 0|\theta) = 1 - \theta$

Probability distribution over X

- $Bern(X|\theta) = \theta^{N_1}(1 - \theta)^{N_0}$
- if the data are iid, the likelihood has the same shape
- there are $N_1 = \sum_{i=1}^{N} \mathbb{1}(X_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{1}(X_i = 0)$ tails
- N_0 and N_1 are called **sufficient statistics** (this is **all** we need to know about data to infer θ)

Bernoulli distribution recap

$$Bern(X|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$
$$N = N_0 + N_1$$

Mean:

- $\mathbb{E}(X) = \theta$

Variance:

- $var(X) = \theta(1-\theta)$

The problem: continuing

Let's demistyfy the examplary problem more:

- suppose the data consists of the count of the number of heads N_1 observed in a fixed number $N = N_1 + N_0$ of trials
- $N_1 \sim Bin(N, \theta)$
- binomial pmf: $Bin(k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$
- term $\binom{n}{k}$ is constant independent of θ, thus the **binomial sampling model is the same as the likelihood for the Bernoulli model**

Likelihood: in general

Likelihood:

- a tool for summarizing the data's evidence about unknown parameter in the model
- as below: considered as a function of θ
- or: is the likelihood function (of θ)
- the probability of "the value x of X for the parameter value θ"

Discrete probability distribution

$$\mathcal{L}(x \mid \theta) = p_\theta(x) = P_\theta(X = x)$$

Continuous probability distribution

$$\mathcal{L}(x \mid \theta) = f_\theta(x)$$

Likelihood of our problem

$$\mathcal{L}(\mathcal{D} \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$$

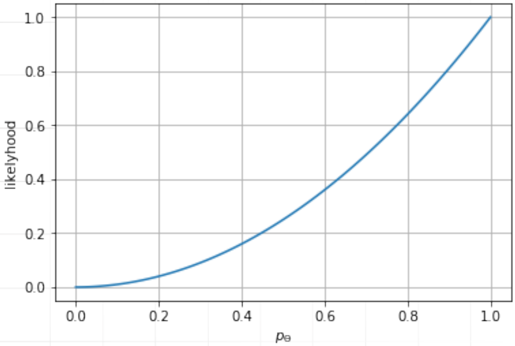


Figure: Likelihood $\mathcal{D}$={HH}

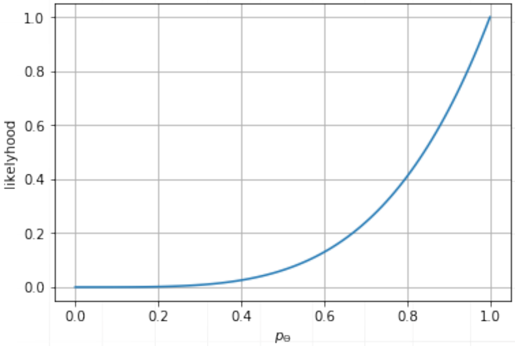


Figure: Likelihood $\mathcal{D}$={HHHH}

What about prior?

- need of prior with support over [0,1] interval
- easier, if it would have the same form as likelyhood, for some prior parameters γ_1 and γ_2:

$$p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_2}$$

- easy evaluation of posterior: adding exponents

$$\mathcal{L}(\theta \mid \mathcal{D})p(\theta) = \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1}(1-\theta)^{\gamma_2} = \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_2}$$

Conjugate priors

When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood.

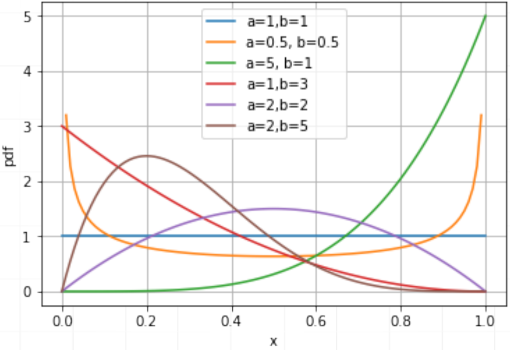
Conjugate priors

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+n-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$p(\tilde{x}=1)=\frac{\alpha'}{\alpha'+\beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x}\mid\alpha',\beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+rn$	$\alpha-1$ total successes, $\beta-1$ failures[note 1] (i.e., $\frac{\beta-1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k+\sum_{i=1}^n x_i,\ \frac{\theta}{n\theta+1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x}\mid k',\theta')$ (negative binomial)
			$\alpha,\ \beta$[note 3]	$\alpha+\sum_{i=1}^n x_i,\ \beta+n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x}\mid\alpha',\frac{1}{1+\beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+(c_1,\ldots,c_k)$, where c_i is the number of observations in category i	α_i-1 occurrences of category i[note 1]	$p(\tilde{x}=i)=\frac{\alpha_i'}{\sum_i\alpha_i'}=\frac{\alpha_i+c_i}{\sum_i\alpha_i+n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+\sum_{i=1}^n \mathbf{x}_i$	α_i-1 occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}}\mid\boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n=N,\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\ \beta$	$\alpha+n,\ \beta+\sum_{i=1}^n x_i-n$	$\alpha-1$ experiments, $\beta-1$ total failures[note 1]	

please check: https://en.wikipedia.org/wiki/Conjugate_prior (source)

Beta distribution

- conjugate prior for the Bernoulli, binomial, negative binomial and geometric distributions
- $Beta(\theta|a, b) \sim \theta^{a-1}(1-\theta)^{b-1}$



Beta distribution

- required $a, b > 0$
- if $a = b = 1$, we get the uniform distirbution
- if a and b are both less than 1, we get a bimodal distribution with “spikes” at 0 and 1
- if a and b are both greater than 1, the distribution is unimodal

Distribution properties

mean=$\frac{a}{a+b}$, mode=$\frac{a-1}{a+b-2}$, var=$\frac{ab}{(a+b)^2(a+b+1)}$

Beta prior

$Beta(\theta|a,b) \sim \theta^{a-1}(1-\theta)^{b-1}$

- prior parameters *a* and *b* are called **hyper-parameters**
- set *a* and *b* to encode your prior belief

Example

- to encode our beliefs that θ has mean 0.7 and standard deviation 0.2, we set a =2.975 and b =1.275
- to encode our beliefs that θ has mean 0.15 and that we think it lives in the interval (0.05, 0.30), we find a =4.5 and b =25.5

Posterior

Multiply the likelihood by the beta prior:

$$p(\theta|\mathcal{D}) \propto Bin(N_1|\theta, N_0 + N_1)Beta(\theta|a, b) \propto Beta(\theta|N_1 + a, N_0 + b)$$

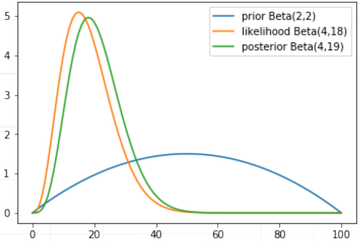


Figure: Beta(2,2) prior updated with Binomial likelihood with sufficient statistics $N_1 = 3, N_0 = 17$ yealds Beta(5,19)

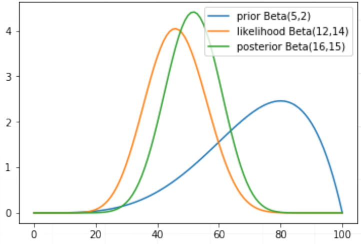


Figure: Beta(5,2) prior updated with Binomial likelihood with sufficient statistics $N_1 = 11, N_0 = 13$ yealds Beta(16,15)

Remark: two ways of updating posterior

Updating the posterior **sequentially** is equivalent to updating in a **single batch**

Let D_a and D_b are two datasets with sufficient statistics N_1^a, N_0^a and N_1^b, N_0^b; let $N_1 = N_1^a + N_1^b$ and $N_0 = N_0^a + N_0^b$ be the sufficient statistics of the combined datasets

- Batch mode:

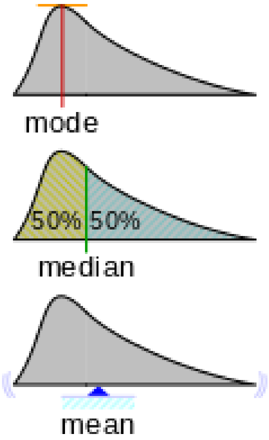
$$p(\theta|\mathcal{D}_a, \mathcal{D}_b) \propto Bin(N_1|\theta, N_1 + N_0) Beta(\theta|a, b) \propto Beta(\theta|N_1 + a, N_0 + b)$$

- Sequential mode:

$$\begin{aligned} p(\theta|\mathcal{D}_a, \mathcal{D}_b) &\propto p(\theta|\mathcal{D}_b) p(\theta|\mathcal{D}_a)) \\ &\propto Bin(N_1^b|\theta, N_1^b + N_0^b) Beta(\theta|N_1^a + a, N_0^a + b) \\ &\propto Beta(\theta|N_1^a + N_1^b + a, N_0^a + N_0^b + b) \end{aligned}$$

REMARK! **Online learning**

Mean, mode, median



Posterior mean and mode

MAP

$$\hat{\theta}_{MAP} = \frac{a + N_1 - 1}{a + b + N - 2}$$

When we use a uniform prior, then the MAP estimate reduces to the MLE, which is just the empirical fraction of heads:

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

Posterior mean

$$\bar{\theta} = \frac{a + N_1}{a + b + N}$$

Posterior predictive distribution

How to make prediction of future observable data?

Predicting the probability of heads in a single future trial under a Beta(a, b) posterior:

$$\begin{aligned} p(\tilde{x} = 1|\mathcal{D}) &= \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D})dx \\ &= \int_0^1 \theta Beta(\theta|a, b)d\theta = \mathbb{E}(\theta|\mathcal{D}) = \frac{a}{a+b} \end{aligned}$$

The mean of the posterior predictive distribution is equivalent (in this case) to plugging in the posterior mean parameters: $p(\tilde{x}|\mathcal{D}) = Ber(\tilde{x}|\mathbb{E}[\theta|\mathcal{D}])$

Overfitting

- let assume $p(\tilde{x}|\mathcal{D}) = Ber(\tilde{x}|\hat{\theta}_{MLE})$
- and $N = 3$ with 3 tails in a row
- MLE is $\hat{\theta} = 0/3 = 0$
- this makes the observed data as probable as possible
- **BUT** we predict that heads are impossible

This is called: **zero count problem** or the **sparse data problem**. Approximation can perform quite poorly.