Machine Learning
CMPT 726
Simon Fraser University

Binomial Parameter Estimation

Outline

- Maximum Likelihood Estimation
- Smoothed Frequencies, Laplace Correction.
- Bayesian Approach.
  - Conjugate Prior.
  - Uniform Prior.

Coin Tossing

- Let's say you're given a coin, and you want to find out $P(heads)$, the probability that if you flip it it lands as "heads".
- Flip it a few times: $H\ H\ T$
- $P(heads) = 2/3$, no need for CMPT726
- Hmm... is this rigorous? Does this make sense?

Coin Tossing

- Let's say you're given a coin, and you want to find out $P(heads)$, the probability that if you flip it it lands as "heads".
- Flip it a few times: $H\ H\ T$
- $P(heads) = 2/3$, no need for CMPT726
- Hmm... is this rigorous? Does this make sense?

Coin Tossing

- Let's say you're given a coin, and you want to find out $P(heads)$, the probability that if you flip it it lands as "heads".
- Flip it a few times: $H\ H\ T$
- $P(heads) = 2/3$, no need for CMPT726
- Hmm... is this rigorous? Does this make sense?

Coin Tossing

- Let's say you're given a coin, and you want to find out $P(heads)$, the probability that if you flip it it lands as "heads".
- Flip it a few times: $H\ H\ T$
- $P(heads) = 2/3$, no need for CMPT726
- Hmm... is this rigorous? Does this make sense?

Coin Tossing - Model

- Bernoulli distribution $P(heads) = \mu$, $P(tails) = 1 - \mu$
- Assume coin flips are independent and identically distributed (i.i.d.)
  - i.e. All are separate samples from the Bernoulli distribution
- Given data $\mathcal{D} = \{x_1, \ldots, x_N\}$, heads: $x_i = 1$, tails: $x_i = 0$, the likelihood of the data is:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

Maximum Likelihood Estimation

- Given $\mathcal{D}$ with h heads and t tails
- What should μ be?
- Maximum Likelihood Estimation (MLE): choose μ which maximizes the likelihood of the data

$$\mu_{ML} = \arg\max_{\mu} p(\mathcal{D}|\mu)$$

- Since $\ln(\cdot)$ is monotone increasing:

$$\mu_{ML} = \arg\max_{\mu} \ln p(\mathcal{D}|\mu)$$

Maximum Likelihood Estimation

- Likelihood:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Log-likelihood:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \ln \mu + (1-x_n)\ln(1-\mu)$$

- Take derivative, set to 0:

$$\frac{d}{d\mu}\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \frac{1}{\mu} - (1-x_n)\frac{1}{1-\mu} = \frac{1}{\mu}h - \frac{1}{1-\mu}t$$

$$\Rightarrow \mu = \frac{h}{t+h}$$

Maximum Likelihood Estimation

- Likelihood:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Log-likelihood:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \ln \mu + (1-x_n)\ln(1-\mu)$$

- Take derivative, set to 0:

$$\frac{d}{d\mu}\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \frac{1}{\mu} - (1-x_n)\frac{1}{1-\mu} = \frac{1}{\mu}h - \frac{1}{1-\mu}t$$

$$\Rightarrow \mu = \frac{h}{t+h}$$

Maximum Likelihood Estimation

- Likelihood:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Log-likelihood:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \ln \mu + (1-x_n)\ln(1-\mu)$$

- Take derivative, set to 0:

$$\frac{d}{d\mu}\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \frac{1}{\mu} - (1-x_n)\frac{1}{1-\mu} = \frac{1}{\mu}h - \frac{1}{1-\mu}t$$

$$\Rightarrow \mu = \frac{h}{t+h}$$

Maximum Likelihood Estimation

- Likelihood:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Log-likelihood:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \ln \mu + (1-x_n)\ln(1-\mu)$$

- Take derivative, set to 0:

$$\frac{d}{d\mu}\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \frac{1}{\mu} - (1-x_n)\frac{1}{1-\mu} = \frac{1}{\mu}h - \frac{1}{1-\mu}t$$

$$\Rightarrow \mu = \frac{h}{t+h}$$

Maximum Likelihood Estimation

- Likelihood:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Log-likelihood:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \ln \mu + (1-x_n)\ln(1-\mu)$$

- Take derivative, set to 0:

$$\frac{d}{d\mu}\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} x_n \frac{1}{\mu} - (1-x_n)\frac{1}{1-\mu} = \frac{1}{\mu}h - \frac{1}{1-\mu}t$$

$$\Rightarrow \mu = \frac{h}{t+h}$$

MLE Estimate: The 0 problem.

- *h* heads, *t* tails, *n = h+t*.
- Practical problems with using the MLE $\frac{h}{n}$

➢ If *h* or *t* are 0, the 0 prob may be multiplied with other nonzero probs (singularity).

➢ If *n* = 0, no estimate at all. This happens quite often in high-dimensional spaces.

Smoothing Frequency Estimates

- *h* heads, *t* tails, *n = h+t*.
- Prior probability estimate *p*.
- Equivalent Sample Size *m*.
- m-estimate = $\frac{h + mp}{n + m}$
- Interpretation: we started with a "virtual" sample of *m* tosses with *mp* heads.
- *P = ½,m=2* ➔ **Laplace correction** = $\frac{h+1}{n+2}$

Bayesian Approach

- Key idea: don't even try to pick specific parameter value μ – use a **probability distribution over parameter values**.
- Learning = use Bayes' theorem to update probability distribution.
- Prediction = **model averaging.**

Prior Distribution over Parameters

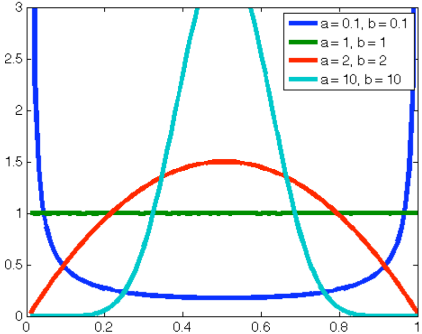
- Could use uniform distribution.
  - Exercise: what does uniform over [0,1] look like?
- What if we don't think prior distribution is uniform?
- Use **conjugate prior**.
  - Prior has parameters *a, b* – "hyperparameters".
  - Prior $P(\mu|a,b) = f(a,b)$ is some function of hyperparameters.
  - Posterior has same functional form $f(a',b')$ where a',b' are updated by Bayes' theorem.

Beta Distribution

- We will use the Beta distribution to express our prior knowledge about coins:

$$Beta(\mu|a,b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{normalization} \mu^{a-1}(1-\mu)^{b-1}$$

- Parameters a and b control the shape of this distribution



Posterior

$$\begin{aligned} P(\mu|\mathcal{D}) &\propto P(\mathcal{D}|\mu)P(\mu) \\ &\propto \underbrace{\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}}_{likelihood} \underbrace{\mu^{a-1}(1-\mu)^{b-1}}_{prior} \\ &\propto \mu^{h}(1-\mu)^{t}\mu^{a-1}(1-\mu)^{b-1} \\ &\propto \mu^{h+a-1}(1-\mu)^{t+b-1} \end{aligned}$$

- Simple form for posterior is due to use of conjugate prior
- Parameters a and b act as extra observations
- Note that as $N = h + t \to \infty$, prior is ignored

Posterior

$$\begin{aligned} P(\mu|\mathcal{D}) &\propto P(\mathcal{D}|\mu)P(\mu) \\ &\propto \underbrace{\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}}_{likelihood} \underbrace{\mu^{a-1}(1-\mu)^{b-1}}_{prior} \\ &\propto \mu^h(1-\mu)^t \mu^{a-1}(1-\mu)^{b-1} \\ &\propto \mu^{h+a-1}(1-\mu)^{t+b-1} \end{aligned}$$

- Simple form for posterior is due to use of conjugate prior
- Parameters a and b act as extra observations
- Note that as $N = h + t \to \infty$, prior is ignored

Posterior

$$
\begin{aligned}
P(\mu|\mathcal{D}) &\propto P(\mathcal{D}|\mu)P(\mu) \\
&\propto \underbrace{\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}}_{likelihood} \underbrace{\mu^{a-1}(1-\mu)^{b-1}}_{prior} \\
&\propto \mu^{h}(1-\mu)^{t}\mu^{a-1}(1-\mu)^{b-1} \\
&\propto \mu^{h+a-1}(1-\mu)^{t+b-1}
\end{aligned}
$$

- Simple form for posterior is due to use of conjugate prior
- Parameters a and b act as extra observations
- Note that as $N = h + t \to \infty$, prior is ignored

Bayesian Point Estimation

- What if a Bayesian **had** to guess a single parameter value given hyperdistribution *P?*
- Use expected value $E_P(\mu)$.
  - E.g., for P = Beta(μ|a,b) we have $E_P(\mu) = a/a+b$.
- If we use uniform prior *P*, what is $E_P(\mu|D)$?
- The Laplace correction!