Bayesian Methods 1

Chris Williams

School of Informatics, University of Edinburgh

September 2014
Overview

- Introduction to Bayesian Statistics: Learning a Bernoulli probability
- Learning a discrete distribution
- Learning the mean of a Gaussian
- Exponential family
- Readings: Murphy §3.3 (Beta), §3.4 (Dirichlet), §4.6.1 (Gaussian)
Bayesian vs Frequentist Inference

**Frequentist**
- Assumes that there is an unknown but fixed parameter $\theta$
- Estimates $\theta$ with some confidence
- Prediction by using the estimated parameter value

**Bayesian**
- Represents uncertainty about the unknown parameter
- Uses probability to quantify this uncertainty. Unknown parameters as random variables
- Prediction follows rules of probability
Frequentist method

- Model $p(x|\theta, M)$, data $D = \{x_1, \ldots, x_N\}$

$$\hat{\theta} = \arg\max_{\theta} p(D|\theta, M)$$

- Prediction for $x_{n+1}$ is based on $p(x_{n+1}|\hat{\theta}, M)$
Bayesian method

- Prior distribution $p(\theta|M)$
- Posterior distribution $p(\theta|D, M)$

$$p(\theta|D, M) = \frac{p(D|\theta, M)p(\theta|M)}{p(D|M)}$$

- Making predictions

$$p(x_{N+1}|D, M) = \int p(x_{N+1}, \theta|D, M) \, d\theta$$

$$= \int p(x_{N+1}|\theta, D, M)p(\theta|D, M) \, d\theta$$

$$= \int p(x_{N+1}|\theta, M)p(\theta|D, M) \, d\theta$$

Interpretation: average of predictions $p(x_{N+1}|\theta, M)$ weighted by $p(\theta|D, M)$

- Marginal likelihood (important for model comparison)

$$p(D|M) = \int P(D|\theta, M)P(\theta|M) \, d\theta$$
Bayes, MAP and Maximum Likelihood

\[ p(x_{N+1}|D, M) = \int p(x_{N+1}|\theta, M)p(\theta|D, M) \, d\theta \]

▶ Maximum a posteriori value of \( \theta \)

\[ \theta_{MAP} = \text{argmax}_\theta \, p(\theta|D, M) \]

Note: not invariant to reparameterization (cf ML estimator)

▶ If posterior is sharply peaked about the most probable value \( \theta_{MAP} \) then

\[ p(x_{N+1}|D, M) \approx p(x_{N+1}|\theta_{MAP}, M) \]

▶ In the limit \( n \to \infty \), \( \theta_{MAP} \) converges to \( \hat{\theta} \) (as long as \( p(\hat{\theta}) \neq 0 \))

▶ Bayesian approach most effective when data is limited, \( N \) is small
Learning probabilities: thumbtack example

Frequentist Approach

- The probability of heads $\theta$ is unknown
- Given iid data, estimate $\theta$ using an estimator with good properties (e.g. ML estimator)
Likelihood

- Likelihood for a sequence of heads (1) and tails (0)

\[ p(1100\ldots001|\theta) = \theta^{N_1}(1 - \theta)^{N_0} \]

- MLE

\[ \hat{\theta} = \frac{N_1}{N_1 + N_0} \]
Learning probabilities: thumbtack example

Bayesian Approach: (a) the prior

- Prior density $p(\theta)$, use Beta distribution

$$p(\theta) = \text{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

for $\alpha, \beta > 0$

- Properties of the Beta distribution

$$E[\theta] = \int \theta p(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$\text{var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$
Examples of the Beta distribution

Beta(0.5,0.5)  Beta(1,1)

Beta(3,2)  Beta(15,10)
Bayesian Approach: (b) the posterior

\[ p(\theta|D) \propto p(\theta)p(D|\theta) \]
\[ \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1} \theta^{N_1} (1 - \theta)^{N_0} \]
\[ \propto \theta^{\alpha+N_1-1} (1 - \theta)^{\beta+N_0-1} \]

- Posterior is also a Beta distribution \( \sim \text{Beta}(\alpha + N_1, \beta + N_0) \)
- The Beta prior is \textit{conjugate} to the binomial likelihood (i.e. prior and posterior have the same parametric form)
- \( \alpha \) and \( \beta \) can be thought of as imaginary counts, with \( \alpha + \beta \) as the equivalent sample size

[cointoss demo]
Bayesian Approach: (c) making predictions

\[ p(X_{N+1} = \text{heads}|D, M) = \int p(X_{N+1} = \text{heads}|\theta)p(\theta|D, M) \, d\theta \]

\[ = \int \theta \text{Beta}(\alpha + N_1, \beta + N_0) \, d\theta \]

\[ = \frac{\alpha + N_1}{\alpha + \beta + N} \]
Beyond Conjugate Priors

The thumbtack came from a magic shop → a mixture prior

\[ p(\theta) = 0.4\text{Beta}(20, 0.5) + 0.2\text{Beta}(2, 2) + 0.4\text{Beta}(0.5, 20) \]
Generalization to multinomial variables

- Dirichlet prior

\[ p(\theta_1, \ldots, \theta_r) = \text{Dir}(\alpha_1, \ldots, \alpha_r) \propto \prod_{i=1}^{r} \theta_i^{\alpha_i-1} \]

with

\[ \sum_i \theta_i = 1, \quad \alpha_i > 0 \]

- \( \alpha_i \)'s are imaginary counts, \( \alpha = \sum_i \alpha_i \) is equivalent sample size

- Properties

\[ E(\theta_i) = \frac{\alpha_i}{\alpha} \]

- Dirichlet distribution is conjugate to the multinomial likelihood
Examples of Dirichlet Distributions

[Source: https://projects.csail.mit.edu/church/wiki/Models_with_Unbounded_Complexity]
likelihood \propto \prod_{i=1}^{r} \theta_i^{N_i}

- Show that MLE $\hat{\theta}_i = \frac{N_i}{N}$

- Posterior distribution

$$p(\theta | N_1, \ldots, N_r) \propto \prod_{i=1}^{r} \theta_i^{\alpha_i + N_i - 1}$$

- Marginal likelihood

$$p(D | M) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{i=1}^{r} \frac{\Gamma(\alpha_i + N_i)}{\Gamma(\alpha_i)}$$
Inferring the mean of a Gaussian

- **Likelihood**
  
  \[ p(x | \mu) \sim N(\mu, \sigma^2) \]

- **Prior**
  
  \[ p(\mu) \sim N(\mu_0, \sigma_0^2) \]

- **Given data** \( D = \{x_1, \ldots, x_N\} \), what is \( p(\mu | D) \)?
\[ p(\mu|D) \sim N(\mu_N, \sigma_N^2) \]

with

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

\[ \mu_N = \frac{n \sigma_0^2}{n \sigma_0^2 + \sigma^2 \bar{x}} + \frac{\sigma^2}{n \sigma_0^2 + \sigma^2} \mu_0 \]

\[ \frac{1}{\sigma_N^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \]

- See Murphy §4.6.1 for details
The exponential family

- Any distribution over some $x$ that can be written as
  \[ P(x|\eta) = h(x)g(\eta) \exp(\eta^T u(x)) \]
  with $h$ and $g$ known, is in the exponential family of distributions.
- Many common distributions are in the exponential family. A notable exception is the $t$-distribution.
- The $\eta$ are called the natural parameters of the distribution.
- For most distributions, the common representation (and parameterization) does not take the exponential family form.
- So sometimes useful to convert to exponential family representation and find the natural parameters.
- Exercise: Why not try this for some of the distributions that we’ve seen already!
Conjugate exponential models

- If the prior takes the same functional form as the posterior for a given likelihood, a prior is said to be *conjugate* for that likelihood.
- There is a conjugate prior for any exponential family distribution.
- If the prior and likelihood are conjugate and exponential, then the model is said to be *conjugate exponential*.
- In conjugate exponential models, the Bayesian integrals can be done analytically.
Reflecting on Conjugacy

- All of the priors that we have seen so far are conjugate
- Good thing: easy to do the sums
- Bad thing: prior distribution should match beliefs. Does a Beta distribution match your beliefs? Is it good enough?
- Certainly not always
- Use of approximate inference methods for non-conjugate models (see later in MLPR)
Comparing Bayesian and Frequentist approaches

- **Frequentist**: fix \( \theta \), consider all possible data sets generated with \( \theta \) fixed
- **Bayesian**: fix \( D \), consider all possible values of \( \theta \)
- One view is that Bayesian and Frequentist approaches have different definitions of what it means to be a good estimator
Summary of Bayesian Methods

- Maximum likelihood fails to capture prior or uncertainty
- Need to use a prior distribution (maximum likelihood equals MAP with uniform prior)
- Prior distribution might have its own parameters (usually called hyper-parameters)
- MAP fails to capture uncertainty, need full posterior distribution
- Prediction using MAP parameters does not capture uncertainty
- Do inference by marginalization. Inference and Learning are just using the rules of probability