Bayesian Methods for Parameter Estimation

Chris Williams

School of Informatics, University of Edinburgh

October 2007

Overview

- Introduction to Bayesian Statistics: Learning a Probability
- Learning the mean of a Gaussian
- Readings: Bishop §2.1 (Beta), §2.2 (Dirichlet), §2.3.6 (Gaussian), Heckerman tutorial section 2

Bayesian vs Frequentist Inference

Frequentist

- Assumes that there is an unknown but fixed parameter θ
- Estimates θ 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 $p(\mathbf{x}|\theta, M)$, data $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$

$$\hat{\theta} = \mathrm{argmax}_\theta \; p(D|\theta, M)$$

- Prediction for $\mathbf{x}_{n+1}$ is based on $p(\mathbf{x}_{n+1}|\hat{\theta}, M)$

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

$$\begin{aligned} p(\mathbf{x}_{n+1}|D, M) &= \int p(\mathbf{x}_{n+1}, \theta|D, M)\ d\theta \\ &= \int p(\mathbf{x}_{n+1}|\theta, D, M)p(\theta|D, M)\ d\theta \\ &= \int p(\mathbf{x}_{n+1}|\theta, M)p(\theta|D, M)\ d\theta \end{aligned}$$

  Interpretation: average of predictions $p(\mathbf{x}_{n+1}|\theta, M)$ weighted by $p(\theta|D, M)$
- Marginal likelihood (important for model comparison)

Bayes, MAP and Maximum Likelihood

$$p(\mathbf{x}_{n+1}|D, M) = \int p(\mathbf{x}_{n+1}|\theta, M)p(\theta|D, M)\ d\theta$$

- *Maximum a posteriori* value of θ

$$\theta_{MAP} = \mathrm{argmax}_\theta\ p(\theta|D, M)$$

  Note: not invariant to reparameterization (cf ML estimator)
- If posterior is sharply peaked about the most probable value θ_{MAP} then

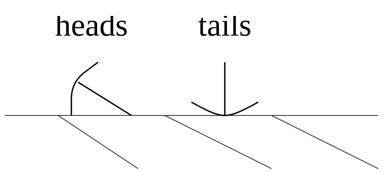
$$p(\mathbf{x}_{n+1}|D, M) \simeq p(\mathbf{x}_{n+1}|\theta_{MAP}, M)$$

- In the limit $n \to \infty$, θ_{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 θ is unknown
- Given iid data, estimate θ using an estimator with good properties (e.g. ML estimator)



Likelihood

- Likelihood for a sequence of heads and tails

$$p(hhth \dots tth|\theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- MLE

$$\hat{\theta} = \frac{n_h}{n_h + n_t}$$

Learning probabilities: thumbtack example

Bayesian Approach: (a) the prior

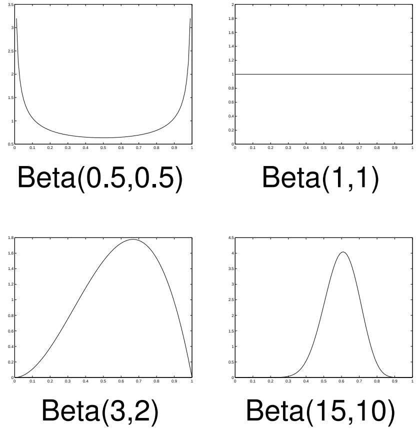
- Prior density $p(\theta)$, use beta distribution

$$p(\theta) = \mathrm{Beta}(\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}$$

  for $\alpha_h, \alpha_t > 0$
- Properties of the beta distribution

$$E[\theta] = \int \theta p(\theta) = \frac{\alpha_h}{\alpha_h + \alpha_t}$$

Examples of the Beta distribution

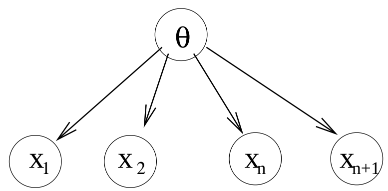


Bayesian Approach: (b) the posterior

$$\begin{aligned} p(\theta|D) &\propto p(\theta)p(D|\theta) \\ &\propto \theta^{\alpha_h - 1}(1-\theta)^{\alpha_t - 1}\theta^{n_h}(1-\theta)^{n_t} \\ &\propto \theta^{\alpha_h + n_h - 1}(1-\theta)^{\alpha_t + n_t - 1} \end{aligned}$$

- Posterior is also a Beta distribution $\sim \mathrm{Beta}(\alpha_h + n_h, \alpha_t + n_t)$
- The Beta prior is *conjugate* to the binomial likelihood (i.e. they have the same parametric form)
- α_h and α_t can be thought of as imaginary counts, with $\alpha = \alpha_h + \alpha_t$ as the equivalent sample size

Bayesian Approach: (c) making predictions



$$\begin{aligned} p(X_{n+1} = heads|D, M) &= \int p(X_{n+1} = heads|\theta)p(\theta|D, M)\, d\theta \\ &= \int \theta\ \mathrm{Beta}((\alpha_h + n_h, \alpha_t + n_t)\, d\theta \\ &= \frac{\alpha_h + n_h}{\alpha + n} \end{aligned}$$

Beyond Conjugate Priors

- The thumbtack came from a magic shop → a mixture prior

$$p(\theta) = 0.4\mathrm{Beta}(20, 0.5) + 0.2\mathrm{Beta}(2, 2) + 0.4\mathrm{Beta}(0.5, 20)$$

Generalization to multinomial variables

- Dirichlet prior

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

  with

$$\sum_i \theta_i = 1, \qquad \alpha_i > 0$$

- α_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

- 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

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\overline{x} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

- See Bishop §2.3.6 for details

Comparing Bayesian and Frequentist approaches

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