Parametric Models: Prior Information, From Models to Answers

Praddep Ravikumar (Instructor), HMW-Alexander (Noter)

January 25, 2017

Back to Index

Contents

Resources

- Lecture

---

1 Bayesian Learning

Given a prior knowledge to estimate the model.

Bayesian Learning:

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

or equivalently

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$$

Likelihood measures the fitness between data and parameters, Prior is the knowledge how possible the parameters to be.

- Prior information encoded as a distribution over possible values of parameter.
- Using the Bayes rule to get an updated posterior distribution over parameters.

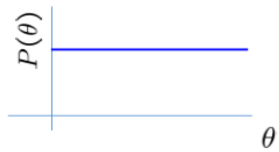
1.1 Prior Distribution

1.1.1 Where to get

- Represents expert knowledge (philosophical approach)
- Simple posterior form (engineer's approach)

1.1.2 Uniformative priors

Simple distribution.



1.1.3 Conjugate Priors

- Closed-form representation of posterior
- prior and posterior have the same algebraic form as a function of parameters

Bernoulli Example: (Binomial's conjugate prior is Beta distribution)

- Likelihood in Bernoulli model: $P(D|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_2}$
- Prior is Beta distribution: $P(\theta) = \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_2-1}}{B(\beta_1,\beta_2)} \sim Beta(\beta_1, \beta_2)$
- Posterior is also Beta distribution: $P(\theta|D) \sim Beta(\beta_1+\alpha_1, \beta_2+\alpha_2)$

Multinomial example: (Multinomial's conjugate prior is Dirichelet distribution)

- Likelihood is Multinomial($\theta = \{\theta_1, \ldots, \theta_k\}$), $P(D|\theta) = \prod_{i=1}^k \theta_i^{\alpha_i}$, $\alpha_i \in \{0,1\}$ is the data D, $\sum_{i=1}^k \theta_i = 1$.
- Prior is Dirichlet distribution: $P(\theta) = \frac{\prod_{i=1}^k \theta_i^{\beta_i-1}}{B(\beta_1,\ldots,\beta_k)} \sim Dirichlet(\beta_1, \ldots, \beta_k)$
- Posterior is also dirichlet distribution: $P(\theta|D) \sim Dirichlet(\beta_1+\alpha_1, \ldots, \beta_k+\alpha_k)$

As we get more samples, effect of prior is "washed out"

2 Maximum A Posteriori Estimation

Choose θ that maximizes a posterior probability: $\hat{\theta}_{MAP} = \arg\max_\theta P(\theta|D)$

$$\begin{aligned} \hat{\theta}_{MAP} &= \arg\max\nolimits_\theta P(\theta|D) \\ &= \arg\max\nolimits_\theta P(D|\theta)P(\theta) \end{aligned} \tag{1}$$

Bernoulli example:

$$\begin{aligned} P(\theta|D) &\sim Beta(\beta_1+\alpha_1, \beta_2+\beta_2) \\ \hat{\theta}_{MAP} &= \frac{\alpha_1+\beta_1-1}{\alpha_1+\beta_1+\alpha_2+\beta_2-2} \end{aligned} \tag{2}$$

2.1 MLE vs. MAP

- MLE: Choose value that maximizes the probability of observed data
- MAP: Choose value that is mot probable given observed data and prior belief
- When prior is a uniform distribution, MLE=MAP.

2.2 MAP for Gaussian mean and variance

- Conjugate priors
  - Gaussian prior
  - Variance: Wishart Distribution