Bayesian Learning

Lecture 6 - Large sample approximations. Classification.

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

LINKÖPING UNIVERSITY

Lecture overview

- **Classification**

- **Naive Bayes**

- **Normal approximation** of posterior

- **Logistic regression** - demo in R

Bayesian classification

- **Classification: output is a discrete label**.
  - ▶ Binary (0-1). Spam/Ham.
  - ▶ Multi-class. ($c = 1, 2, ..., C$). Brand choice.
- **Bayesian classification**

$$\underset{c \in \mathcal{C}}{\operatorname{argmax}} \, p(c|\mathbf{x})$$

where $\mathbf{x} = (x_1, ..., x_p)$ is a covariate/feature vector.

- **Discriminative models** - model $p(c|\mathbf{x})$ directly.
  - ▶ Examples: logistic regression, support vector machines.
- **Generative models** - Use Bayes' theorem

$$p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$$

with class-conditional distribution $p(\mathbf{x}|c)$ and prior $p(c)$.

  - ▶ Examples: discriminant analysis, naive Bayes.

Naive Bayes

- By Bayes' theorem

$$p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$$

- $p(c)$ can be estimated by Multinomial-Dirichlet analysis.

- $p(\mathbf{x}|c)$ can be $N(\theta_c, \Sigma_c)$

- $p(\mathbf{x}|c)$ can be very high-dimensional and hard to estimate.

- Even with binary features, the outcome space of $p(\mathbf{x}|c)$ can be huge.

- Naive Bayes: features are assumed independent

$$p(\mathbf{x}|c) = \prod_{j=1}^{n} p(x_j|c)$$

Classification with logistic regression

- Response is assumed to be **binary** ($y = 0$ or 1).
- Example: Spam/Ham. Covariates: \$-symbols, etc.
- **Logistic regression**

$$\Pr(y_i = 1 \mid x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

- **Likelihood**

$$p(\mathbf{y}|\mathbf{X}, \beta) = \prod_{i=1}^{n} \frac{[\exp(x_i'\beta)]^{y_i}}{1 + \exp(x_i'\beta)}.$$

- Prior $\beta \sim N(0, \tau^2 I)$. Posterior is non-standard (demo later).
- Alternative: **Probit regression**

$$Pr(y_i = 1|x_i) = \Phi(x_i'\beta)$$

- **Multi-class** ($c = 1, 2, ..., C$) logistic regression

$$\Pr(y_i = c \mid x_i) = \frac{\exp(x_i'\beta_c)}{\sum_{k=1}^{C} \exp(x_i'\beta_k)}$$

Large sample approximate posterior

■ **Taylor expansion of log-posterior** around mode $\theta = \tilde{\theta}$:

$$\ln p(\theta|\mathbf{y}) = \ln p(\tilde{\theta}|\mathbf{y}) + \frac{\partial \ln p(\theta|\mathbf{y})}{\partial \theta}|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta})$$

$$+ \frac{1}{2!} \frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial \theta^2}|_{\theta=\tilde{\theta}}(\theta - \tilde{\theta})^2 + ...$$

■ From the definition of the posterior mode:

$$\frac{\partial \ln p(\theta|\mathbf{y})}{\partial \theta}|_{\theta=\tilde{\theta}} = 0$$

■ So, in **large samples** (higher order terms negligible):

$$p(\theta|\mathbf{y}) \approx p(\tilde{\theta}|\mathbf{y}) \exp\left(-\frac{1}{2} J_{\mathbf{y}}(\tilde{\theta})(\theta - \tilde{\theta})^2\right)$$

where $J_{\mathbf{y}}(\tilde{\theta}) = -\frac{\partial^2 \ln p(\theta|\mathbf{y})}{\partial \theta^2}|_{\theta=\tilde{\theta}}$ is the **observed information**.

■ **Approximate normal posterior** in large samples.

$$\theta|\mathbf{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$$

Example: gamma posterior

- **Poisson model**: $\theta | y_1, ..., y_n \sim Gamma(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$

  $\log p(\theta | y_1, ..., y_n) \propto (\alpha + \sum_{i=1}^{n} y_i - 1) \log \theta - \theta(\beta + n)$

- First derivative of log density

  $$\frac{\partial \ln p(\theta | \mathbf{y})}{\partial \theta} = \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\theta} - (\beta + n)$$

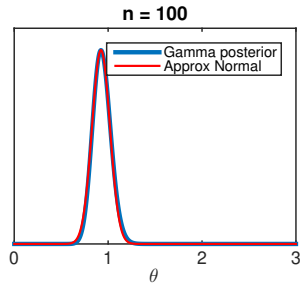 $$\tilde{\theta} = \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}$$
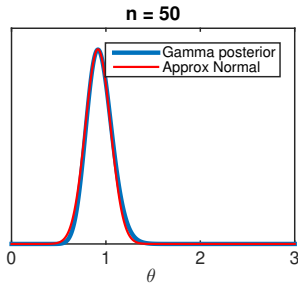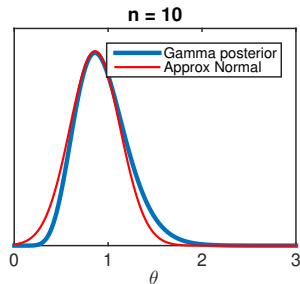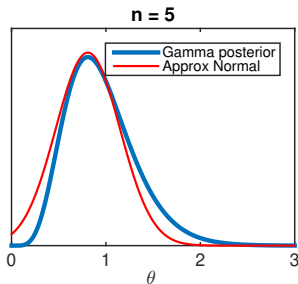
- Second derivative at mode $\tilde{\theta}$

  $$\frac{\partial^2 \ln p(\theta | \mathbf{y})}{\partial \theta^2} \big|_{\theta = \tilde{\theta}} = -\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\left(\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n} \right)^2} = -\frac{(\beta + n)^2}{\alpha + \sum_{i=1}^{n} y_i - 1}$$

- **Normal approximation**

  $$N \left[\frac{\alpha + \sum_{i=1}^{n} y_i - 1}{\beta + n}, \frac{\alpha + \sum_{i=1}^{n} y_i - 1}{(\beta + n)^2} \right]$$

Example: gamma posterior

Normal approximation of posterior

- $\theta|\mathbf{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_y^{-1}(\tilde{\theta})\right]$ works also when θ is a vector.

- How to compute $\tilde{\theta}$ and $J_{\mathbf{y}}(\tilde{\theta})$?

- Standard **optimization routines** may be used. (`optim.r`).
  - ▶ **Input**: expression proportional to $\log p(\theta|\mathbf{y})$. Initial values.
  - ▶ **Output**: $\log p(\tilde{\theta}|\mathbf{y})$, $\tilde{\theta}$ and Hessian matrix $(-J_{\mathbf{y}}(\tilde{\theta}))$.

- **Automatic differentiation** (autodiff in Python, ForwardDiff in Julia, R?)

- **Re-parametrization** may improve normal approximation. [Don't forget the **Jacobian**!]
  - ▶ If $\theta \geq 0$ use $\phi = \log(\theta)$.
  - ▶ If $0 \leq \theta \leq 1$, use $\phi = \ln[\theta/(1-\theta)]$.

- **Heavy tailed approximation**: $\theta|\mathbf{y} \overset{approx}{\sim} t_v\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ for suitable degrees of freedom v.

Reparametrization - Gamma posterior

- Poisson model. Reparameterize to $\phi = \log(\theta)$.
- Change-of-variables formula from a basic probability course

$$\log p(\phi|y_1, ..., y_n) \propto \left(\alpha + \sum_{i=1}^{n} y_i - 1\right)\phi - \exp(\phi)(\beta + n) + \phi$$

- Taking first and second derivatives and evaluating at $\tilde{\phi}$ gives

$$\tilde{\phi} = \log\left(\frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}\right) \text{ and } \frac{\partial^2 \ln p(\phi|y)}{\partial \phi^2}\Big|_{\phi=\tilde{\phi}} = \alpha + \sum_{i=1}^{n} y_i$$
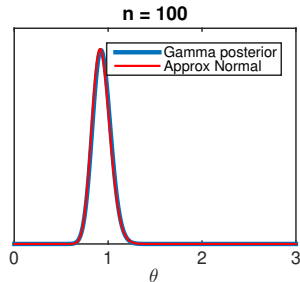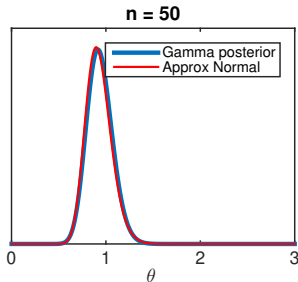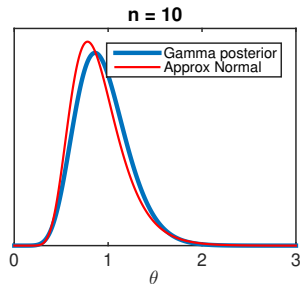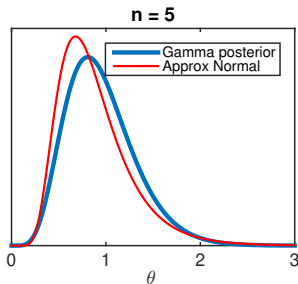
- So, the normal approximation for $p(\phi|y_1, ...y_n)$ is

$$\phi = \log(\theta) \sim N\left[\log\left(\frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}\right), \frac{1}{\alpha + \sum_{i=1}^{n} y_i}\right]$$

which means that $p(\theta|y_1, ...y_n)$ is log-normal:

$$\theta|\mathbf{y} \sim LN\left[\log\left(\frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}\right), \frac{1}{\alpha + \sum_{i=1}^{n} y_i}\right]$$

Reparametrization – Gamma posterior

Normal approximation of posterior

- Even if the posterior of θ is approx normal, **interesting functions** of $g(\theta)$ may not be (e.g. predictions).

- But approximate posterior of $g(\theta)$ can be obtained by **simulating** from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.

- Posterior of **Gini coefficient**
  - Model: $x_1, ..., x_n | \mu, \sigma^2 \sim LN(\mu, \sigma^2)$.
  - Let $\phi = \log(\sigma^2)$. And $\theta = (\mu, \phi)$.
  - Joint posterior $p(\mu, \phi)$ may be approximately normal: $\theta | \mathbf{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
  - Simulate $\theta^{(1)}, ..., \theta^{(N)}$ from $N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$.
  - Compute $\sigma^{(1)}, ..., \sigma^{(N)}$.
  - Compute $G^{(i)} = 2\Phi\left(\sigma^{(i)}/\sqrt{2}\right)$ for $i = 1, ..., N$.