

BAYESIAN LEARNING - LECTURE 6

Mattias Villani

**Division of Statistics
Department of Computer and Information Science
Linköping University**

LECTURE OVERVIEW

- ▶ Classification
- ▶ Naive Bayes
- ▶ Normal approximation of posterior
- ▶ Logistic regression - demo in R

BAYESIAN CLASSIFICATION

- ▶ **Classification: output is a discrete label.** Examples:
 - ▶ binary (0-1). Spam/Ham.
 - ▶ Multi-class. ($c = 1, 2, \dots, C$). $\{iPhone, Android, Windows, Other\}$.
- ▶ **Bayesian classification**

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

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

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

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

and model 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 $N(\theta_c, \Sigma_c)$ or mixture of normals (see last module).
- ▶ $p(\mathbf{x}|c)$ 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

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

- ▶ Naive Bayes solution

$$p(c|\mathbf{x}) \propto \left[\prod_{j=1}^n p(x_j|c) \right] p(c)$$

CLASSIFICATION WITH LOGISTIC REGRESSION

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

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

- ▶ Likelihood

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

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

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

- ▶ **Multi-class** ($c = 1, 2, \dots, C$) 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 the posterior mode $\theta = \tilde{\theta}$:

$$\begin{aligned}\ln p(\theta|y) &= \ln p(\tilde{\theta}|y) + \frac{\partial \ln p(\theta|y)}{\partial \theta} \Big|_{\theta=\tilde{\theta}} (\theta - \tilde{\theta}) \\ &\quad + \frac{1}{2!} \frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2} \Big|_{\theta=\tilde{\theta}} (\theta - \tilde{\theta})^2 + \dots\end{aligned}$$

- From the definition of the posterior mode:

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

- So, in **large samples** (where we can ignore higher order terms):

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

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

- **Approximate posterior**

$$\theta|y \stackrel{approx}{\sim} N[\tilde{\theta}, J_y^{-1}(\tilde{\theta})]$$

EXAMPLE: GAMMA POSTERIOR

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

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

- First derivative of log density

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

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

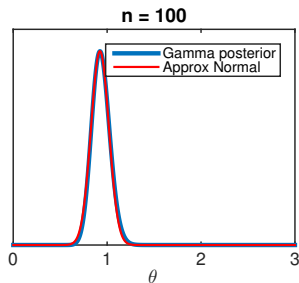
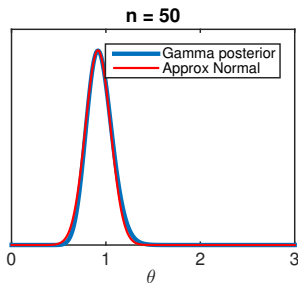
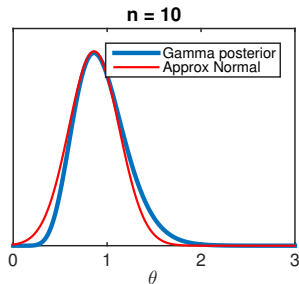
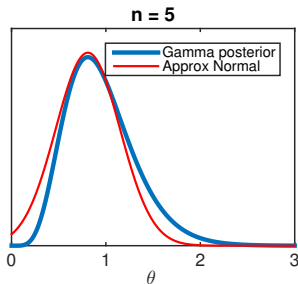
- Second derivative at mode $\tilde{\theta}$

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

- So, the normal approximation is

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

EXAMPLE: GAMMA POSTERIOR



NORMAL APPROXIMATION OF POSTERIOR

- ▶ $\theta|y \stackrel{approx}{\sim} N[\tilde{\theta}, J_y^{-1}(\tilde{\theta})]$ works also when θ is a vector.
- ▶ How to compute $\tilde{\theta}$ and $J_y(\tilde{\theta})$?
- ▶ Standard **optimization routines** may be used. (optim.r).
 - ▶ **Input**: an expression proportional to $\log p(\theta|y)$ and initial values.
 - ▶ **Output**: $\log p(\tilde{\theta}|y)$, $\tilde{\theta}$ and Hessian matrix $(-J_y(\tilde{\theta}))$.
- ▶ **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**: $\theta|y \stackrel{approx}{\sim} t_\nu[\tilde{\theta}, J_y^{-1}(\tilde{\theta})]$ for suitable degrees of freedom ν .

EXAMPLE: GAMMA POSTERIOR - REPARAM.

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

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

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

$$\tilde{\phi} = \log \left(\frac{\alpha + \sum_{i=1}^n y_i}{\beta + n} \right) \text{ and } \frac{\partial^2 \ln p(\phi|y)}{\partial \phi^2} \Big|_{\phi=\tilde{\phi}} = \alpha + \sum_{i=1}^n y_i$$

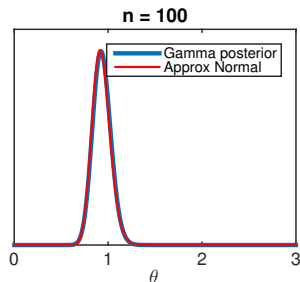
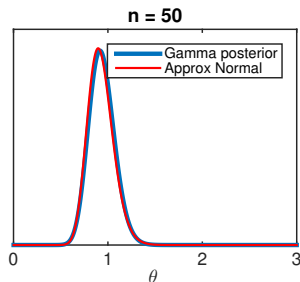
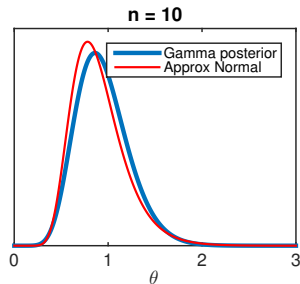
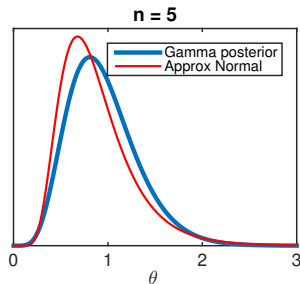
- ▶ So, the normal approximation for $p(\phi|y_1, \dots, y_n)$ is

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

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

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

EXAMPLE: GAMMA POSTERIOR - REPARAMETERIZED



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