Bayesian Learning - Lecture 6

Mattias Villani

Division of Statistics
Department of Computer and Information Science
Linköping University

Lecture overview

- Classification
- Naive Bayes
- Logistic regression
- Normal approximation of posterior

Bayesian classification

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

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

  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

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

  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 $\beta \sim N(0, \lambda I)$. Posterior is non-standard.
- Alternative: **Probit regression** (see Lab 3)

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

- **Multi-class** ($c = 0, 1, 2, ..., C$) logistic regression

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

Large sample approximate posterior

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

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

- From the definition of the posterior mode:

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

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

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

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

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

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

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

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

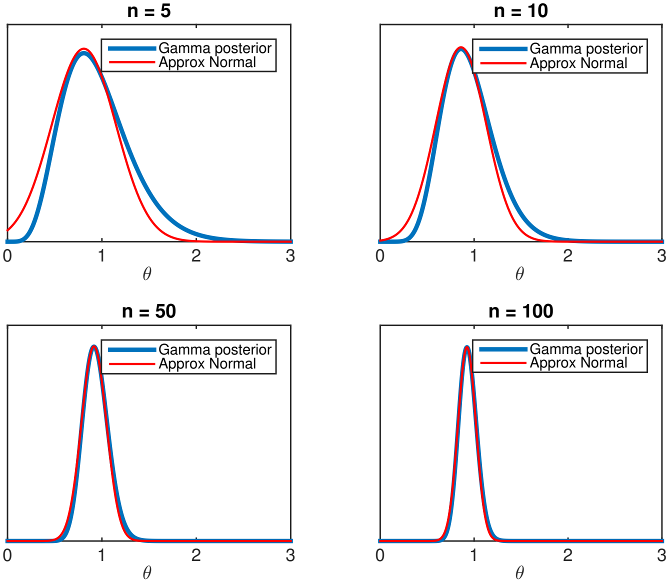
- Second derivative at mode $\tilde{\theta}$

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

- 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

- $\theta|y \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ 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**: an expression proportional to $\log p(\theta|y)$ and initial values.
  - **Output**: $p(\tilde{\theta}|y)$, $\tilde{\theta}$ and Hessian matrix $(-J_{\mathbf{y}}(\tilde{\theta}))$.
- **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**: $\theta|y \overset{approx}{\sim} t_v\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ for suitable degrees of freedom v.

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

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