Bayesian Learning

Lecture 6 - Classification. Large sample approximations.

Bertil Wegmann

Department of Computer and Information Science
Linköping University

LINKÖPING UNIVERSITY

Lecture overview

- Classification
- Logistic regression
- Normal approximation to the posterior

Bayesian classification

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

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

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

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

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

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

  - Example: naive Bayes.

Classification with logistic regression

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

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

- **Likelihood**

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

- Prior $\beta \sim N\left(0, \frac{1}{\lambda} I\right)$. Posterior is non-standard.
- Alternative: **Probit regression**

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

- **Multi-class** ($c = 1, 2, ..., C$) logistic regression ($\beta_1 = 0$)

$$\Pr(y_i = c \mid x_i, \beta_1, \ldots, \beta_C) = \frac{\exp(x_i'\beta_c)}{\sum_{k=1}^{C} \exp(x_i'\beta_k)}$$

Large sample approximate posterior

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

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

- From the definition of the posterior mode: $\frac{\partial \ln p(\theta|\mathrm{y})}{\partial \theta}|_{\theta=\tilde{\theta}} = 0$
- So, in **large samples** (higher order terms negligible):

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

  where $J(\tilde{\theta}) = -\frac{\partial^2 \ln p(\theta|\mathrm{y})}{\partial \theta^2}|_{\theta=\tilde{\theta}}$ is the **observed information** at the mode $\tilde{\theta}$.
- **Approximate normal posterior** in large samples.

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

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

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

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

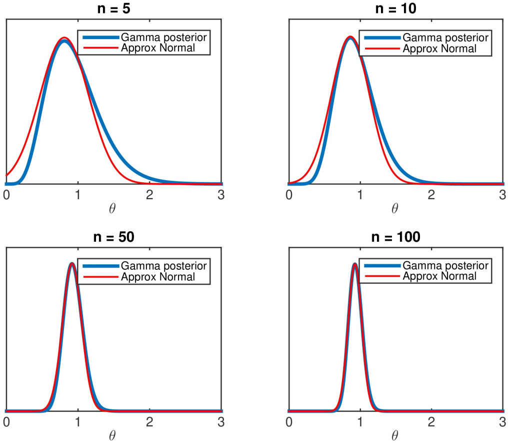
- Second derivative at the mode $\tilde{\theta}$

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

- **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 to the posterior

- $\theta|\mathrm{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_y^{-1}(\tilde{\theta})\right]$ works also when θ is a vector.
- How to compute $\tilde{\theta}$ and $J_{\mathrm{y}}(\tilde{\theta})$?
- Standard **optimization routines** may be used. (`optim.r`).
  - **Input**: expression proportional to $\log p(\theta|\mathrm{y})$. Initial values.
  - **Output**: $\log p(\tilde{\theta}|\mathrm{y})$, $\tilde{\theta}$ and Hessian matrix $(-J_{\mathrm{y}}(\tilde{\theta}))$.
- **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**: $\theta|\mathrm{y} \overset{approx}{\sim} t_v\left[\tilde{\theta}, J_{\mathrm{y}}^{-1}(\tilde{\theta})\right]$ 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

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

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

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

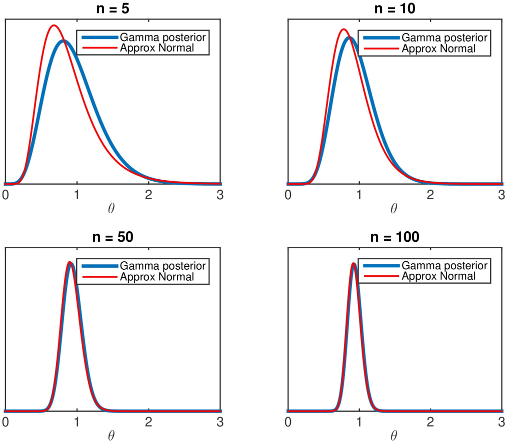
- So, the normal approximation for $p(\phi|y_1, ...y_n)$ is

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

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

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

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

- Posterior of **Gini coefficient**
  - Model: $y_1, ..., y_n|\mu, \sigma^2 \sim LN(\mu, \sigma^2)$.
  - Let $\phi = \log(\sigma^2)$. And $\theta = (\mu, \phi)$.
  - Joint posterior $p(\mu, \phi|\mathrm{y})$ may be approximately normal: $\theta|\mathrm{y} \overset{approx}{\sim} N\left[\tilde{\theta}, J_{\mathrm{y}}^{-1}(\tilde{\theta})\right]$.
  - Simulate $\theta^{(1)}, ..., \theta^{(N)}$ from $N\left[\tilde{\theta}, J_{\mathrm{y}}^{-1}(\tilde{\theta})\right]$.
  - Compute $\sigma^{(1)}, ..., \sigma^{(N)}$.
  - Compute $G^{(j)} = 2\Phi\left(\sigma^{(j)}/\sqrt{2}\right) - 1$ for $j = 1, ..., N$.