Bayesian Statistics I

Lecture 6 - Large sample approximations. Classification.

Mattias Villani

Department of Statistics
Stockholm University

Department of Computer and Information Science
Linköping University

mattiasvillani.com @matvil mattiasvillani

Lecture overview

- **Classification**
- **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**

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

where $\mathrm{x} = (x_1, ..., x_p)^\top$ 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 $p(c)$.

  - Examples: discriminant analysis, naive Bayes.

Naive Bayes

- By Bayes' theorem

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

- $p(c)$ can be estimated by Multinomial-Dirichlet analysis.
- $p(\mathrm{x}|c)$ can be $N(\theta_c, \Sigma_c)$
- $p(\mathrm{x}|c)$ can be very high-dimensional and hard to estimate.
- Even with binary features (e.g. `hasWord('money')` for spam), the outcome space of $p(\mathrm{x}|c)$ can be huge.
- **Naive Bayes**: features are assumed **independent**

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

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**

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

- Prior $\beta \sim N(0, \tau^2 I)$. 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

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

Likelihood asymptotics

- **Taylor expansion of log-likelihood** around the MLE $\theta = \hat{\theta}$:

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

- Higher order terms $(\ldots)$ negligible in large samples.
- From the definition of the MLE:

$$\frac{\partial \ln p(\mathrm{x}|\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = 0$$

- So, in **large samples**

$$p(\mathrm{x}|\theta) \approx p(\mathrm{x}|\hat{\theta}) \exp\left(-\frac{1}{2}J_{\mathrm{x}}(\hat{\theta})(\theta - \hat{\theta})^2\right)$$

- **Observed information**

$$J_{\mathrm{x}}(\hat{\theta}) = -\frac{\partial^2 \ln p(\mathrm{x}|\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}}$$

Likelihood asymptotics

- $J_x(\hat{\theta})$ varies from sample to sample. **Fisher information**

$$I(\theta) = \mathbb{E}_{x|\theta}\left(J_x(\hat{\theta})\right)$$

- Multiparameter **observed information matrix**

$$J_{\boldsymbol{\theta},x}(\hat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

- Example: $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$

$$\frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} = \begin{pmatrix} \frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial\theta_1^2} & \frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} \\ \frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial\theta_1\partial\theta_2} & \frac{\partial^2 \ln p(x|\boldsymbol{\theta})}{\partial\theta_2^2} \end{pmatrix}.$$

Posterior asymptotics

- We can do the same Taylor approximation on log posterior

$$\log p(\boldsymbol{\theta}|\mathrm{x}) = \log p(\mathrm{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathrm{x})$$

- **Approximate normal posterior** in large samples

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

- $\tilde{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{\theta}|\mathrm{x})$ is the posterior mode and
- $J_{\mathrm{x}}^{-1}(\tilde{\boldsymbol{\theta}})$ is now with respect to posterior $\log p(\boldsymbol{\theta}|\mathrm{x})$.
- Likelihood will dominate the prior in large samples so
  - $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$
  - $J_{\mathrm{x}}^{-1}(\tilde{\boldsymbol{\theta}})$ will be close to the **observed information**.
- Important: sufficient with proportional form

$$\log p(\boldsymbol{\theta}|\mathrm{x}) = \log p(\mathrm{x}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

.

Example: gamma posterior

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

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

- First derivative of log density

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

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

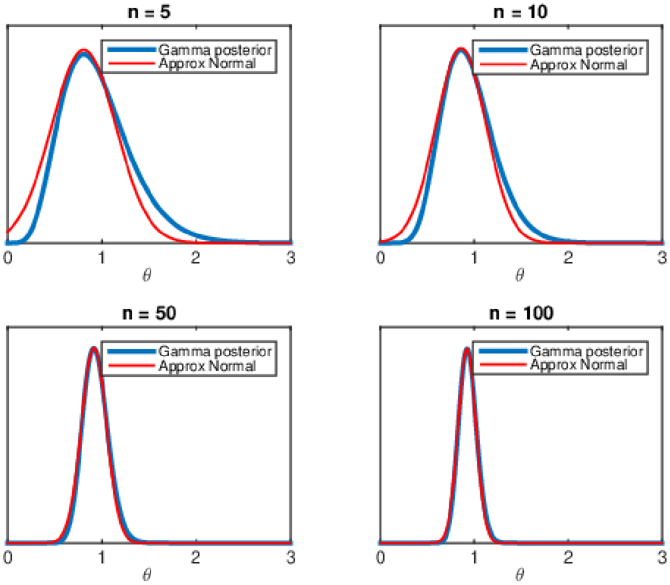
- Second derivative at mode $\tilde{\theta}$

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

- **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 of posterior

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

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

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

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

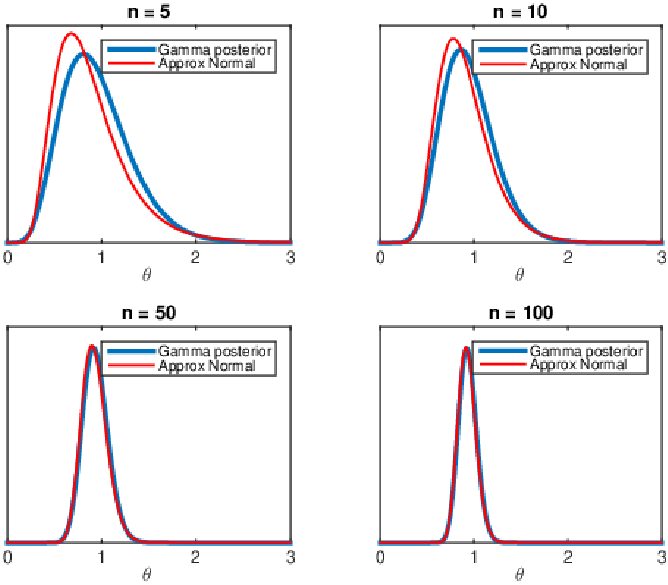
- So, the normal approximation for $p(\phi|y_1, ...y_n)$ is

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

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

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

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

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