

# Bayesian Learning

## Lecture 6 - Classification. Large sample approximations.

Bertil Wegmann

Department of Computer and Information Science  
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. ( $c = 1, 2, \dots, C$ ). Brand choice.

## ■ Bayesian classification

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

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

## ■ Discriminative models - model $p(c|x)$ directly.

- ▶ Examples: logistic regression, support vector machines.

## ■ Generative models - Use Bayes' theorem

$$p(c|x) \propto p(x|c)p(c)$$

with class-conditional distribution  $p(x|c)$  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**

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

- **Likelihood**

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

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

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

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

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

# Large sample approximate posterior

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

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

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

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

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

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

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

## 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

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

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

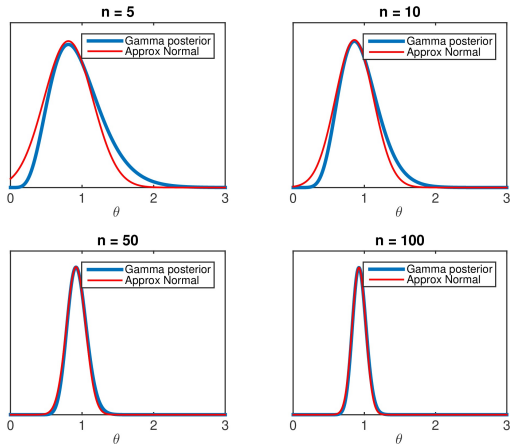
- Second derivative at the mode  $\tilde{\theta}$

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

- **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

- $\theta|y \stackrel{approx}{\sim} N[\tilde{\theta}, J_y^{-1}(\tilde{\theta})]$  works also when  $\theta$  is a vector.
- How to compute  $\tilde{\theta}$  and  $J_y(\tilde{\theta})$ ?
- Standard **optimization routines** may be used. (optim.r).
  - ▶ **Input:** expression proportional to  $\log p(\theta|y)$ . 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:**  $\theta|y \stackrel{approx}{\sim} t_v[\tilde{\theta}, J_y^{-1}(\tilde{\theta})]$  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

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

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

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

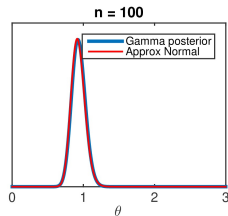
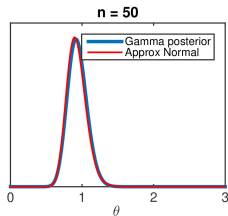
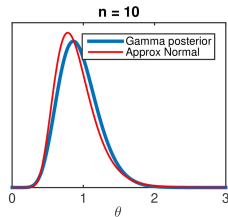
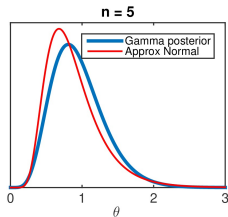
- So, the normal approximation for  $p(\phi|y_1, \dots, y_n)$  is

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

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

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

# Reparametrization - Gamma posterior



# Normal approximation of posterior

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