

Bayesian Learning

Lecture 5 - Large sample approximations. Classification.

Mattias Villani 🧑

Department of Statistics
Stockholm University



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. ($c = 1, 2, \dots, C$). Brand choice.

■ Bayesian classification

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

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

■ Discriminative models - model $p(c|\mathbf{x})$ directly.

- ▶ Examples: logistic regression, support vector machines.

■ Generative models - Use Bayes' theorem

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

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

- ▶ Examples: discriminant analysis, naive Bayes.

Classification with logistic regression

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

$$\Pr(y_i = 1 | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

- **Likelihood**

$$p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n \frac{[\exp(\mathbf{x}_i^\top \boldsymbol{\beta})]^{y_i}}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

- Prior $\boldsymbol{\beta} \sim N(0, \tau^2 I)$. Posterior is non-standard (demo later).
- Alternative: **Probit regression**

$$\Pr(y_i = 1 | \mathbf{x}_i) = \Phi(\mathbf{x}_i^\top \boldsymbol{\beta})$$

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

$$\Pr(y_i = c | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c)}{\sum_{k=1}^C \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_k)}$$

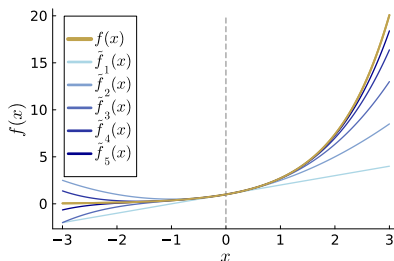
Taylor approximation

- **Taylor approximation** of the function $f(x)$ around $x = a$

$$f(x) \approx f(a) + \sum_{k=0}^K \frac{f^{(k)}(a)}{k!} (x - a)^k$$

- Taylor approximation of $f(x) = \exp(x)$

$$\exp(x) \approx \sum_{k=0}^K \frac{x^k}{k!}$$



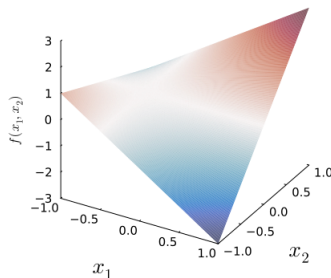
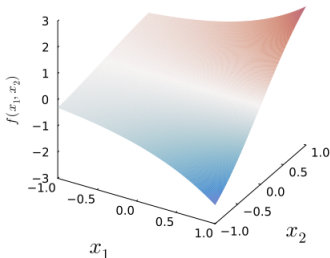
Multi-dimensional Taylor approximation

■ Multi-dimensional Taylor approximation of $f(\mathbf{x})$

$$f(\mathbf{x}) \approx f(\mathbf{a}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} (\mathbf{x} - \mathbf{a}) + \frac{1}{2} (\mathbf{x} - \mathbf{a})^\top \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^\top} (\mathbf{x} - \mathbf{a}) + \dots$$

$$f(x_1, x_2) = \exp(x_1) \sin(x_2)$$

Taylor 2nd order



Likelihood asymptotics

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

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

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

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

- So, in **large samples**

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

- **Observed information**

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

Likelihood asymptotics

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

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

- Multiparameter **observed information matrix**

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

- Example: $\theta = (\theta_1, \theta_2)^T$

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

Posterior asymptotics

- We can do the same Taylor approximation on log posterior

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

- **Approximate normal posterior** in large samples

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

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

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

Large sample asymptotics

Normal posterior approximation

The posterior can in large samples be approximated by

$$\boldsymbol{\theta}|\mathbf{y} \stackrel{a}{\sim} \mathcal{N}\left(\tilde{\boldsymbol{\theta}}, J_{\boldsymbol{\theta},\mathbf{y}}^{-1}(\tilde{\boldsymbol{\theta}})\right)$$

where $\tilde{\boldsymbol{\theta}}$ is the posterior mode and

$$J_{\tilde{\boldsymbol{\theta}},\mathbf{y}} = -\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

is the $d \times d$ observed information matrix at $\tilde{\boldsymbol{\theta}}$.

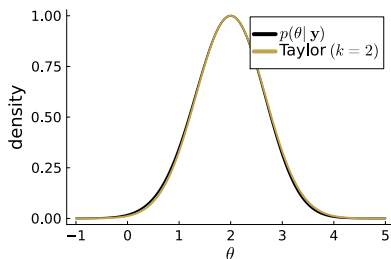
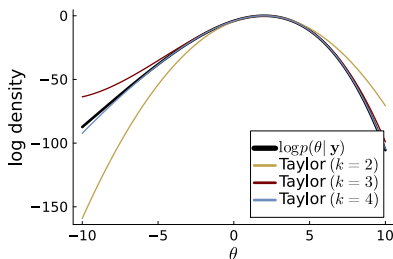
Normal approximation example

Posterior

$$p(\theta|\mathbf{y}) \propto \exp\left(-\exp(\theta/\kappa_0)(\theta - \bar{y})^2\right)$$

where κ_0 is a prior hyperparameter and \bar{y} is the sample mean.

Taylor expansion of log posterior



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

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

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

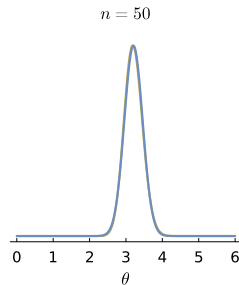
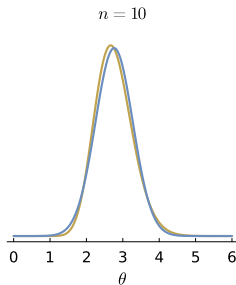
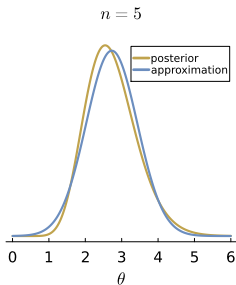
- Second derivative at mode $\tilde{\theta}$

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

- **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 for eBay bidders data



Normal approximation of posterior

- $\theta|\mathbf{y} \stackrel{\text{approx}}{\sim} N\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ 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:** expression proportional to $\log p(\theta|\mathbf{y})$. Initial values.
 - ▶ **Output:** $\log p(\tilde{\theta}|\mathbf{y})$, $\tilde{\theta}$ and Hessian matrix $(-J_{\mathbf{y}}(\tilde{\theta}))$.
- **Automatic differentiation** - 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:** $\theta|\mathbf{y} \stackrel{\text{approx}}{\sim} t_{\nu}\left[\tilde{\theta}, J_{\mathbf{y}}^{-1}(\tilde{\theta})\right]$ for suitable degrees of freedom ν .

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

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

- 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:

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

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 $N\left[\tilde{\theta}, J_y^{-1}(\tilde{\theta})\right]$.
- 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:
 $\theta | y \stackrel{\text{approx}}{\sim} N\left[\tilde{\theta}, J_y^{-1}(\tilde{\theta})\right]$.
 - ▶ Simulate $\theta^{(1)}, \dots, \theta^{(N)}$ from $N\left[\tilde{\theta}, J_y^{-1}(\tilde{\theta})\right]$.
 - ▶ Compute $\sigma^{(1)}, \dots, \sigma^{(N)}$.
 - ▶ Compute $G^{(i)} = 2\Phi\left(\sigma^{(i)} / \sqrt{2}\right)$ for $i = 1, \dots, N$.

Bayesian logistic regression

■ Logistic regression

$$\Pr(y_i = 1 | \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

■ Odds

$$\text{Odds}(\mathbf{x}) \equiv \frac{\Pr(y = 1 | \mathbf{x})}{\Pr(y = 0 | \mathbf{x})} = \exp(\mathbf{x}^\top \boldsymbol{\beta}).$$

■ Odds ratio

$$OR_j = \frac{\text{Odds}(x_1, \dots, x_j + 1, \dots, x_p)}{\text{Odds}(x_1, \dots, x_j, \dots, x_p)} = \exp(\beta_j)$$

■ Likelihood

$$p(\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}) = \prod_{i=1}^n \frac{[\exp(\mathbf{x}_i^\top \boldsymbol{\beta})]^{y_i}}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.$$

■ Normal approximation:

$$\boldsymbol{\beta} | \mathbf{y} \sim N\left(\tilde{\boldsymbol{\beta}}, J_{\mathbf{y}}^{-1}(\tilde{\boldsymbol{\beta}})\right).$$

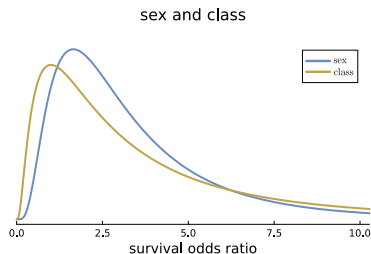
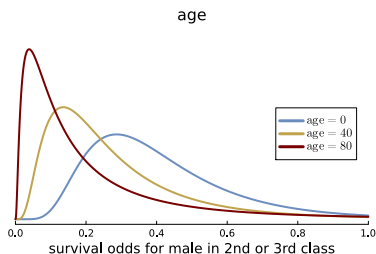
Logistic regression - who survived the Titanic?

■ Prior

$$\beta \sim N(\mu, \Omega)$$

with

$$\mu = (-1, -1/80, 1, 1)^\top \quad \Omega = \begin{pmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 1/(80^2) & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

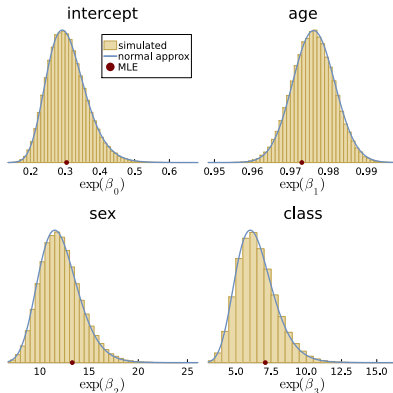


Logistic regression - who survived the Titanic?

Normal posterior approximation

$$\beta | \mathbf{y} \sim N\left(\tilde{\beta}, J_{\mathbf{y}}^{-1}(\tilde{\beta})\right).$$

- Means that the posterior of each β_j is univariate normal.
- Marginal posterior for each $\exp(\beta_j)$ is **LogNormal**.



Logistic regression - who survived the Titanic?

- Comparison with non-informative prior $\beta \sim N(\mathbf{0}, 10^2 I_p)$.

