Bayesian Learning

Lecture 5 - Large sample approximations. Classification.

Mattias Villani 🧔

Department of Statistics
Stockholm University

mattiasvillani.com   @matvil   @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**

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

where $\mathbf{x} = (x_1, ..., x_p)^\top$ 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**

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

- **Likelihood**

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

- 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

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

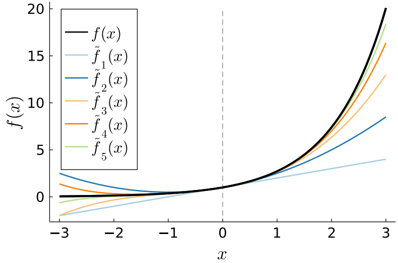
Taylor approximation

- **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 $f(x) = \exp(x)$

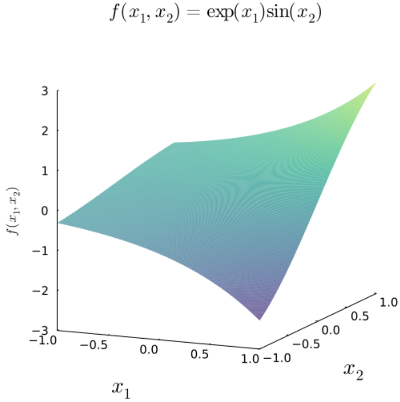
$$\exp(x) \approx \sum_{k=0}^{K} \frac{x^k}{k!}$$

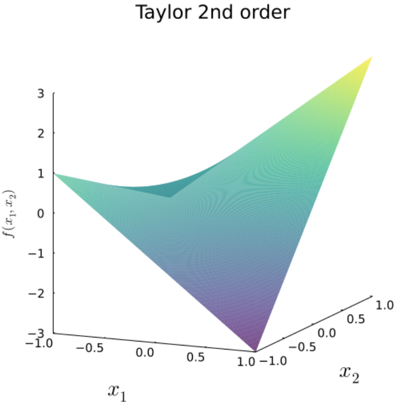


Multi-dimensional Taylor approximation

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

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





Likelihood asymptotics

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

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

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

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

- 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**

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

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**

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

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

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

Posterior asymptotics

- We can do the same Taylor approximation on log posterior

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

- **Approximate normal posterior** in large samples

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

- $\tilde{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{\theta}|\mathbf{x})$ 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

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

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

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

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

Large sample asymptotics

Theorem 2 (large sample normality of posterior). *The posterior distribution of θ conditional on data $\mathbf{y} = (y_1, \ldots, y_n)$ converges to a normal distribution in large samples:*

$$J_{\theta,\mathbf{y}}^{1/2}(\tilde{\theta})(\theta - \tilde{\theta}) \mid \mathbf{y} \xrightarrow{d} \mathrm{N}(0,1), \ \textit{as } n \to \infty,$$

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

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

is the observed information at $\tilde{\theta}$.

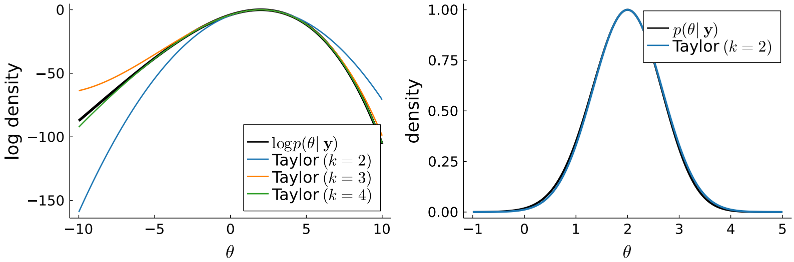
Normal approximation example

- Posterior

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

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

- Taylor expansion of log posterior



Example: gamma posterior

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

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

- First derivative of log density

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

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

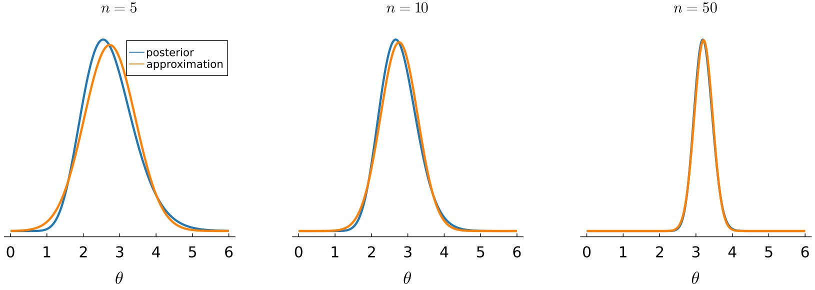
- Second derivative at mode $\tilde{\theta}$

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

- **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

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

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

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

- 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 $i = 1, ..., N$.

Bayesian logistic regression

- **Logistic regression**

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

- **Likelihood**

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

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

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

- Demo time!