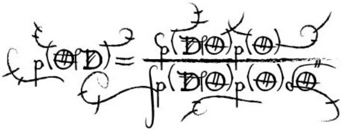


Bayesian Learning 732A46: Lecture 6

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

April 2016

Lecture overview

- Large sample theory
- Classification
- Naive Bayes (generative)
- Logistic regression (discriminative)

The likelihood dominates the prior

A statement I made during the first lecture

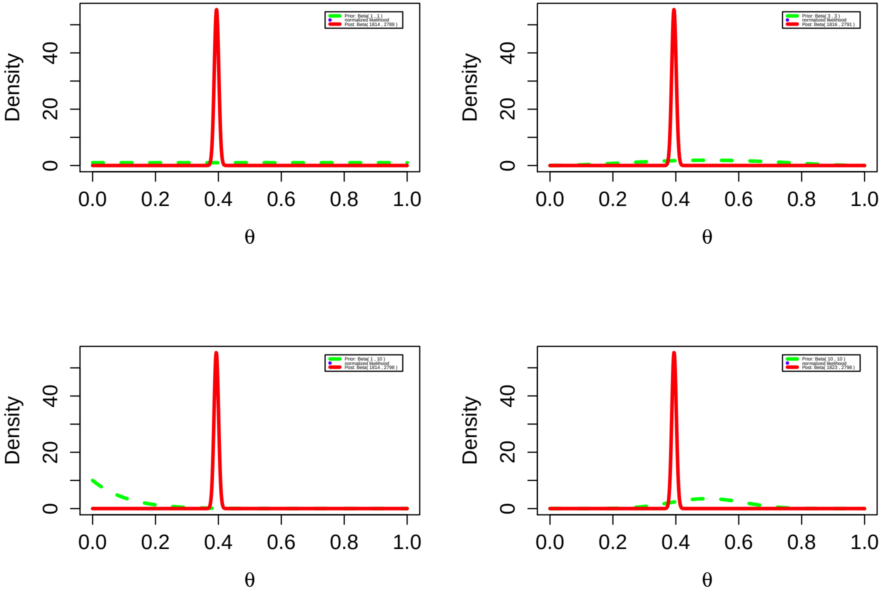
The influence of the prior **vanishes** as more data is collected. In other words, **in large samples the likelihood dominates the prior**. Any reasonable prior results in **essentially the same inferences**.

- Recall the **spam data example**:

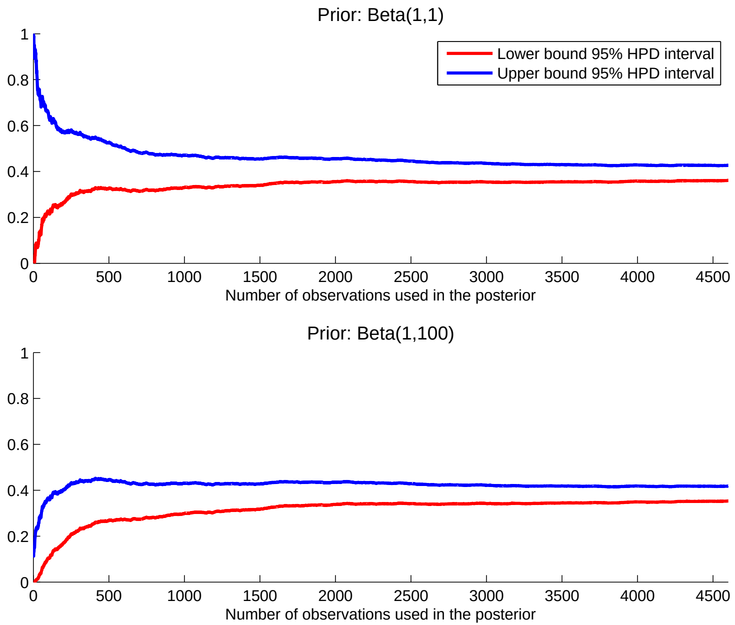
  George has gone through his collection of 4601 e-mails. He classified 1813 of them to be spam (and 2788 non-spam).

- Four different priors gave the **same result**.

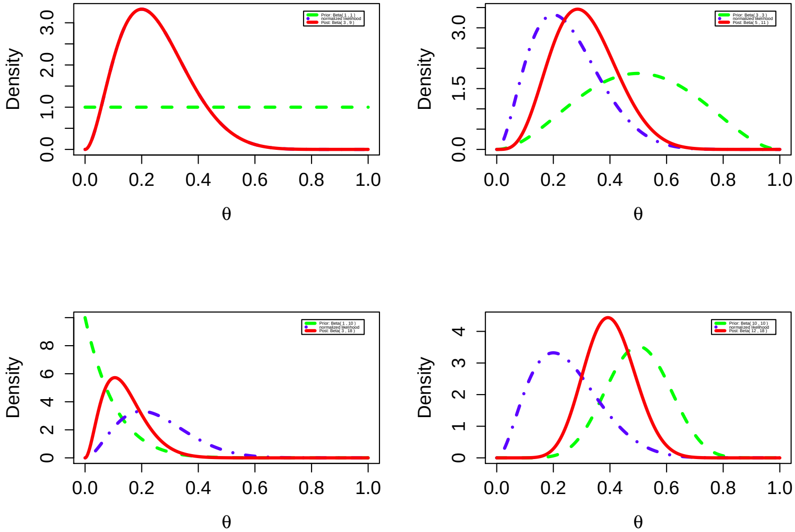
The likelihood dominates the prior, cont.



The likelihood dominates the prior, cont.

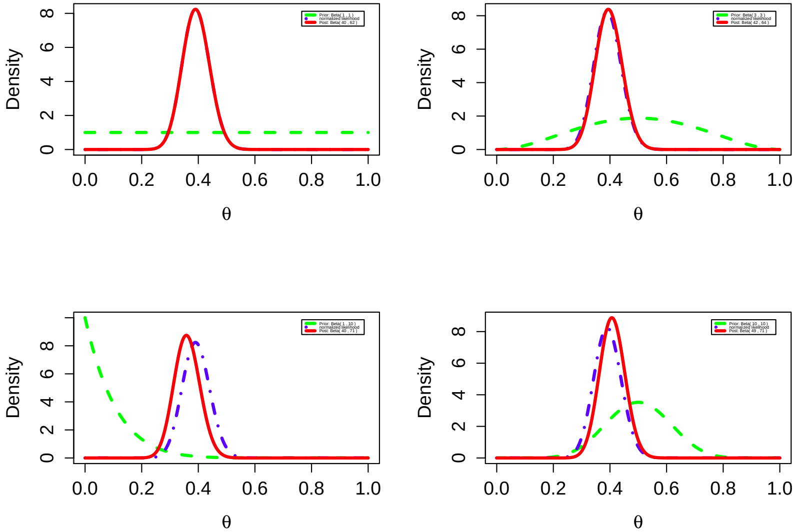


- **Recall:** In small samples it was **far from** a normal distribution...

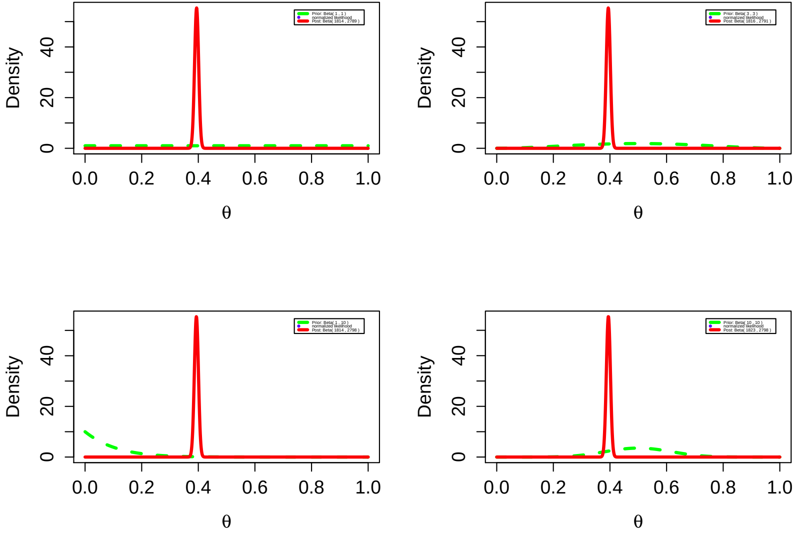


The behaviour of the posterior as the sample size increases

- ... as **more data** entered the estimation, normality becomes more reasonable...



- ... and with all data (a **large** sample)... Note that it also **concentrates**...



Formalizing the statements

- We have made **three observations** as n **increases**
  1. The likelihood **dominates the prior**
  2. The posterior **approaches a normal distribution**
  3. The posterior **concentrates around a value**

2. Taylor series expansion $p(\theta|y)$ w.r.t $\theta \in \mathbb{R}^p$ around the posterior mode $\theta^\star$

$$\log p(\theta|y) = \log p(\theta^\star|y) + \nabla_\theta \log p(\theta^\star|y)'(\theta - \theta^\star) + \frac{1}{2!}(\theta - \theta^\star)'\nabla\nabla'_\theta \log p(\theta^\star|y)(\theta - \theta^\star) + \ldots$$

where

$$\nabla_\theta \log p(\theta^\star|y) = \left.\frac{\partial \ln p(\theta|y)}{\partial \theta}\right|_{\theta=\theta^\star} \in \mathbb{R}^p \quad \textbf{(gradient)}$$

$$\nabla\nabla'_\theta \log p(\theta^\star|y) = \left.\frac{\partial^2 \ln p(\theta|y)}{\partial \theta \partial \theta'}\right|_{\theta=\theta^\star} \in \mathbb{R}^{p\times p} \quad \textbf{(Hessian)}$$

Formalizing statement nbr 2.

- Define the **observed information**

$$J_y(\theta^\star) = -\nabla\nabla'_\theta \log p(\theta^\star|y) \quad (\textbf{negative Hessian})$$

- At the mode **the gradient** is zero (necessary condition to be a mode)

$$\nabla_\theta \log p(\theta^\star|y) = \mathbf{0}$$

- The Taylor series..

$$\log p(\theta|y) = \log p(\theta^\star|y) - \frac{1}{2!}(\theta - \theta^\star)' J_y(\theta^\star)(\theta - \theta^\star) + \ldots$$

- ... can be truncated (in **large samples**)

$$\log p(\theta|y) \approx \log p(\theta^\star|y) - \frac{1}{2!}(\theta - \theta^\star)' J_y(\theta^\star)(\theta - \theta^\star)$$

- ... which is a **quadratic form** in θ... $p(\theta|y)$ is $\propto$ a normal density!

Formalizing statement nbr 2., cont.

- Taking exponents

$$p(\theta|y) \approx \underbrace{p(\theta^\star|y)}_{c} \exp\left(-\frac{1}{2}(\theta-\theta^\star)' J_y(\theta^\star)(\theta-\theta^\star)\right) \propto \mathcal{N}_p(\theta^\star, J_y^{-1}(\theta^\star)).$$

- **Why is this useful?**

  We can approximate the posterior of (many) complex (non-conjugate) models by a normal distribution...

  ... but note that we require the posterior mode and the Hessian **evaluated at the mode**...

  ... can easily be obtained with numerical optimization (e.g. `optim` in `R`).

- **But be aware**

  Posterior might be **multi-modal**.

  Rate of convergence depends on p. Large p will require **very large** n.

Formalizing statement nbr 1.

1. **Recall**: The likelihood dominates the prior.

- Assume conditionally iid. observations

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \quad \text{and} \quad \ell(\theta) = \log p(y|\theta) = \sum_{i=1}^{n} \log p(y_i|\theta) = \sum_{i=1}^{n} \ell_i(\theta).$$

- **Bayes' theorem** in log scale $\log p(\theta|y) = \log p(y|\theta) + \log p(\theta) + c$
- Since

$$\nabla\nabla_\theta' \log p(\theta^\star|y) = \nabla\nabla_\theta' \ell(\theta^\star) + \nabla\nabla_\theta' \log p(\theta^\star)$$

the observed information is

$$J_y(\theta^\star) = \left(-\sum_{i=1}^{n} J_{y_i}(\theta^\star) \right) - J(\theta^\star), \;\; J_{y_i} = -\nabla\nabla_\theta' \ell_i(\theta) \quad J(\theta^\star) = -\nabla\nabla_\theta' \log p(\theta^\star)$$

3. As n increases the curvature is **dominated** by the information part coming from the likelihood. **Recall**

$$\Sigma_n = J_y^{-1}(\theta^\star) \quad \text{[The posterior covariance]}$$

.

Formalizing statement nbr 3.

- **Recall**: The posterior concentrates around a value.
- **Posterior consistency**: the posterior degenerates to the "**true value**".

$$\text{"A value"} = \text{"the true value"}$$

- **But what is the "true value"?**
  - *Mathematical idealization*.
    Let Θ denote the parameter space. The value θ_0 is the "true value" in the sense that the data

$$y \sim f(y) = p(y|\theta_0) \quad \text{for some } \theta_0 \in \Theta.$$

  - **Note**: the data is random (θ_0 is a fixed constant).
- **Proof**: similar but replacing $J_i(\theta^\star)$ by the expected information

$$I_i(\theta_0) = -E_{y_i}\left(\nabla\nabla'_\theta \log p(y_i|\theta_0)\right) \quad \text{so that } I(\theta_0) = \sum_{i=1}^{n} I_i(\theta_0)$$

is the **Fisher information**.

Example: Normal approximation of a gamma posterior

- **Poisson model**: $\theta|y_1,...,y_n \sim \mathrm{Gamma}(\alpha_0 + \sum_{i=1}^n y_i, \beta_0 + n)$

$$\log p(\theta|y_1,...,y_n) \propto (\alpha_0 + \sum_{i=1}^n y_i - 1)\log(\theta) - \theta(\beta_0 + n)$$

- First derivative (**gradient** for $p = 1$) of log density

$$\frac{\partial \ln p(\theta|y)}{\partial \theta} = \frac{\alpha_0 + \sum_{i=1}^n y_i - 1}{\theta} - (\beta_0 + n) = 0 \implies \theta^\star = \frac{\alpha_0 + \sum_{i=1}^n y_i - 1}{\beta_0 + n}$$

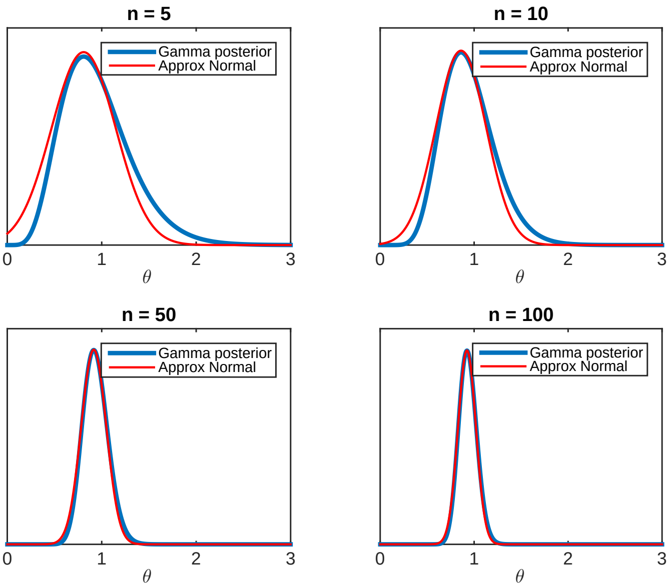
- Second derivative (**Hessian** for $p = 1$) at mode $\theta^\star$

$$\left.\frac{\partial^2 \ln p(\theta|y)}{\partial \theta^2}\right|_{\theta=\theta^\star} = -\frac{\alpha_0 + \sum_{i=1}^n y_i - 1}{\left(\frac{\alpha_0 + \sum_{i=1}^n y_i - 1}{\beta_0 + n}\right)^2} = -\frac{(\beta_0 + n)^2}{\alpha_0 + \sum_{i=1}^n y_i - 1}.$$

- The **normal approximation** is

$$\mathcal{N}\left(\frac{\alpha_0 + \sum_{i=1}^n y_i - 1}{\beta_0 + n}, \frac{\alpha_0 + \sum_{i=1}^n y_i - 1}{(\beta_0 + n)^2}\right).$$

Example: Normal approximation of a gamma posterior, cont.



Normal approximation of posterior

- For complex models / high dimensional θ use standard **optimization routines** (e.g. `optim` in `R`).
  - **Input**: an expression **proportional to** $\log p(\theta|y)$ and initial values.
  - **Output**: $\log p(\theta^\star|y)$, $\theta^\star$ and Hessian matrix $[-J_y(\theta^\star)]$.

- **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 $\phi = \ln[\theta/(1-\theta)]$.

- **Recall change of variables**: Let $p_\theta(\theta)$ be continuous and let $\phi = h(\theta)$ be a one-to-one transform.

$$p_\phi(\phi) = p_\theta(h^{-1}(\phi))|J|, \quad |J| = \text{determinant of } h^{-1}(\phi) \left[1-\mathrm{dim}: \frac{d}{d\phi}h^{-1}(\phi)\right].$$

- Even if $p(\theta|y) \approx \mathcal{N}$, $g(\theta)$ may have a **very complex** posterior...
- ... The **joy of simulating** to the rescue!

Bayesian classification

- **Classification** is like **regression** but with a discrete label as output.
- **Examples**.
  - binary (0-1). Spam/Ham.
  - Multi-class. ($c = 1, 2, ..., C$). $\{iPhone, Android, Windows, Other\}$.
- Let $x = (x_1, ..., x_p)'$ be a vector of p **covariates/features** (inputs).
- **Posterior distribution** over the classes (output)

$$\Pr(c = k|x), \quad k = 1, \ldots, C \quad [= p(c|x)].$$

- The **Bayesian** classifies

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

- **Two approaches**
  1. **Discriminative models** - model $p(c|x)$ directly (logistic regression, SVM)
  2. **Generative models** - Use Bayes' theorem $p(c|x) \propto p(x|c)p(c)$ and model
     (i) the **class-conditional distribution** $p(x|c)$
     (ii) the **prior** $p(c)$.

     Examples: discriminant analysis, naive Bayes.

Generative model: Naive Bayes

- By **Bayes' theorem**

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

- **Example:** Let $c = \{\text{male}, \text{female}\}$ and $x = \{\text{weight}, \text{length}, \text{shoe size}\}$
- **Data:** $\{c_i, x_i\}_{i=1}^n$
- $p(c)$ can be estimated by a conjugate **Bernoulli-Beta** model with data $c_i, i = 1, \ldots, n$ (or **Multinomial-Dirichlet** if $C > 2$)
- **Non-naive**: $p(x|c)$ can be $\mathcal{N}_p(\theta_c, \Sigma_c)$ (or more flexibly a **Finite mixture of normals**, see next module) (**Note**: $p = 3$ in Ex.).
- **Naive Bayes**: features are **assumed independent**

$$p(x|c) = \prod_{j=1}^{p} p(x_j|c) \implies p(c|x) \propto \left[\prod_{j=1}^{p} p(x_j|c)\right] p(c),$$

  Note: the **class-conditionals** are modeled separately.
- **Classify** using probabilities, e.g. $x = \{52 \text{ kg}, 160 \text{ cm}, 36\}$

$$\Pr(\text{female}|x) \propto p(x_1 = 52|\text{female})p(x_2 = 160|\text{female})p(x_3 = 36|\text{female})\Pr(\text{female})$$

Generative model: Naive Bayes, cont.

- The **Naive Bayes** is not necessary a realistic model.
- **Our example**: variables are probably **dependent** (even if gender is known)
- Why don't **always** go **non-Naive**?
  1. Feature vector x might be **very** high-dimensional.
  2. Even with binary features the sample space of x can be **huge**.

Discriminative model: logistic regression

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

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

- **Likelihood**

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

  Note: implicitly conditioning on covariates (they are not modeled)
- **Our example here**: $x_i = \{\text{weight}, \text{length}, \text{shoe size}\}$, for the ith obs.

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

  Note: **not** modeling weight/length/shoe size as in the generative model.
- **Prior** $\beta \sim N(0, \lambda^{-1} I)$ (simple shrinkage prior). **Posterior** is **non-standard**.

Discriminative model: logistic regression, cont.

- Markov Chain Monte Carlo (MCMC) can be used to simulate $p(\beta|y)$.
- We can alternatively obtain a **normal approximation** of $p(\beta|y)$.
- **Homework**: Go through `MainOptimizeSpam.R`.
  - **Learn how to master** the function `optim`
  - **Learn how to code** an expression of the log posterior (only $\propto$ required)
  - **Add a step** where you (given the output from `optim`) simulate from the posterior. **Hint**:

$$p(\beta|y) \approx \mathcal{N}(\theta^\star, \Sigma_\star) \quad \left[\theta^\star\text{: mode}, \Sigma_\star\text{: -Hess}^{-1}(\theta^\star)\right].$$

- Generalization to **multi-class** $(c = 1, \ldots, C)$ **logistic regression**

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