T-61.5140 Machine Learning: Advanced Probabilistic Methods
Pekka Marttinen, Sami Remes (Spring 2015)
Exercise solutions, round 7, Tuesday, 10th March, 2015
https://noppa.aalto.fi/noppa/kurssi/t-61.5140

Problem 1. *"Bayes factors."*

(a) *Suppose we have two bags, each containing a large number of black and white marbles. To learn about the contents of the bags, we have done 5 draws from each bag. After each draw, the marble drawn has been returned to the bag. The draws from the first bag are as follows $x_1 = (B, W, W, B, B)$ and the draws from the second bag are $x_2 = (B, B, B, B, W)$, where $X_{ij} = B$ corresponds to a Black marble and $X_{ij} = W$ to a White marble. Consider two models*

$$M_1 : \textit{the proportions of white marbles are the same in the two bags}$$
$$M_2 : \textit{the proportions of white marbles are different in the two bags.}$$

Assuming that a priori all proportions are equally probable, compute the Bayes factor in favor of M_1. (Hint: Beta distribution is the conjugate prior for the Binomial/Bernoulli likelihood, and a uniform proportion corresponds to the $Beta(1, 1)$ distribution.)

(b) *The same as (a), but now x_1 contains 300 black and 200 white draws, and x_2 250 black and 250 white draws.*

(c) *The same as (a), but now in addition to black and white marbles, some marbles may be red (R) and the observations are as follows: $x_1 = (B, B, W, B, R, R, W, R)$ and $x_2 = (R, B, B, B, B, R, R, B)$. (Hint: Dirichlet distribution is the conjugate prior for the multinomial/categorical likelihood.)*

Solution. **(a, b)** We can express the model M_1 as

$$\begin{aligned} x_i &\sim Binomial(n, p) \quad i = 1, 2 \\ p &\sim Beta(1, 1) \end{aligned}$$

and model M_2 as

$$\begin{aligned} x_i &\sim Binomial(n, p_i) \quad i = 1, 2 \\ p_i &\sim Beta(1, 1) \end{aligned}$$

where x_1 and x_2 denote the number of white marbles drawn from the first and second bag, respectively, and n the total number of marbles in one draw.

For computing the Bayes factor we need the marginal likelihoods for each model, derived as follows.

$$
\begin{aligned}
p(x|M_1) &= \int_0^1 Binomial(x_1|n,p)Binomial(x_2|n,p)Beta(p|1,1)dp \\
&= \binom{n}{x_1}\binom{n}{x_2}\int_0^1 p^{x_1}(1-p)^{n-x_1}p^{x_2}(1-p)^{n-x_2}dp \\
&= \binom{n}{x_1}\binom{n}{x_2}\int_0^1 p^{x_1+x_2+1-1}(1-p)^{2n-x_1-x_2+1-1}dp \\
&= \binom{n}{x_1}\binom{n}{x_2}B(x_1+x_2+1, 2n-x_1-x_2+1).
\end{aligned}
$$

$$
\begin{aligned}
p(x|M_2) &= \int_0^1\int_0^1 Binomial(x_1|n,p_1)Binomial(x_2|n,p_2)Beta(p_1|1,1)Beta(p_2|1,1)dp_1dp_2 \\
&= \binom{n}{x_1}\binom{n}{x_2}\int_0^1\int_0^1 p_1^{x_1}(1-p_1)^{n-x_1}p_2^{x_2}(1-p_2)^{n-x_2}dp_1dp_2 \\
&= \binom{n}{x_1}\binom{n}{x_2}B(x_1+1, n-x_1+1)B(x_2+1, n-x_2+1).
\end{aligned}
$$

So the Bayes factor is given as

$$
K = \frac{p(x|M_1)}{p(x|M_2)} = \frac{B(x_1+x_2+1, 2n-x_1-x_2+1)}{B(x_1+1, n-x_1+1)B(x_2+1, n-x_2+1)}.
$$

In (a) the values were $n = 5$, $x_1 = 2$ and $x_2 = 1$, so the Bayes factor is $K \approx 1.3636$. In (b) $n = 500$, $x_1 = 200$ and $x_2 = 250$, so $K \approx 0.0816$.

(c) Here we modify the model M_1 as

$$
\begin{aligned}
x_i &\sim Multinomial(n,p) \quad i = 1,2 \\
p &\sim Dirichlet(\alpha)
\end{aligned}
$$

and model M_2 as

$$
\begin{aligned}
x_i &\sim Multinomial(n,p_i) \quad i = 1,2 \\
p_i &\sim Dirichlet(\alpha)
\end{aligned}
$$

where x_1 and x_2 denote the counts of each color of marbles drawn from the first and second bag, respectively, and n the total number of marbles in one draw. p is a probability vector denoting the proportions of each color in a bag. Uniform proportions here also correspond to the prior $Dirichlet(\alpha)$ with $\alpha = \mathbf{1} = (1,1,1)^T$.

Marginal likelihoods are again

$$
\begin{aligned}
p(x|M_1) &= \int Multinomial(x_1|p) Multinomial(x_2|p) Dirichlet(p|\alpha) dp \\
&= \frac{n!}{\prod_{k=1}^{3} x_{1k}!} \frac{n!}{\prod_{k=1}^{3} x_{2k}!} \int \prod_{k=1}^{3} p_k^{x_{1k}} \prod_{k=1}^{3} p_k^{x_{2k}} B(\alpha)^{-1} \prod_{k=1}^{3} p_k^{\alpha-1} dp \\
&= \frac{n!}{\prod_{k=1}^{3} x_{1k}!} \frac{n!}{\prod_{k=1}^{3} x_{2k}!} B(\alpha)^{-1} \int \prod_{k=1}^{3} p_k^{x_{1k}+x_{2k}+\alpha-1} dp \\
&= \frac{n!}{\prod_{k=1}^{3} x_{1k}!} \frac{n!}{\prod_{k=1}^{3} x_{2k}!} B(\alpha)^{-1} B(x_1 + x_2 + \alpha)
\end{aligned}
$$

and

$$
\begin{aligned}
p(x|M_2) &= \int Multinomial(x_1|p_1) Multinomial(x_2|p_2) Dirichlet(p_1|\alpha) Dirichlet(p_2|\alpha) dp \\
&= \frac{n!}{\prod_{k=1}^{3} x_{1k}!} \frac{n!}{\prod_{k=1}^{3} x_{2k}!} B(\alpha)^{-2} B(x_1 + \alpha) B(x_2 + \alpha)
\end{aligned}
$$

where $B(\cdot)$ is the multivariate Beta function (normalizing constant of the Dirichlet distribution). So the Bayes factor is given by

$$
K = B(\alpha) \frac{B(x_1 + x_2 + \alpha)}{B(x_1 + \alpha) B(x_2 + \alpha)} \approx 1.1518,
$$

with $x_1 = (3, 2, 3)$, $x_2 = (5, 0, 3)$ (counts of black, white and red marbles, respectively), and $n = 8$.

Problem 2. *"Laplace and BIC approximations for Bayes factors."*

Derive and compute the Bayes factor in Problem 1 (a) and (b) using the Laplace approximation and the BIC approximation. Comment on the accuracy of the approximations. (Optional, not required for the exercise point: do this also for (c).)

Solution. Laplace approximation is given by

$$
L(M) = \log p(x|M) \approx \log p(x, \hat{\theta}|M) + \frac{k}{2} \log(2\pi) - \frac{1}{2} \log |H_{\hat{\theta}}|
$$

where $\hat{\theta}$ is the MAP estimate of the model M parameters and $H_{\hat{\theta}}$ the Hessian (or second derivative) of $-\log p(x, \theta|M)$ at evaluated $\hat{\theta}$, and k is the number of paramters in the model.

BIC is given by

$$
BIC(M) = \log p(x, \hat{\theta}|M) - \frac{k}{2} \log N
$$

where N is the size of the data.

MAP with a uniform prior is equal to the ML estimate, so for the Binomial distributions we have

$$
\hat{p} = \frac{x_1 + x_2}{2n}
$$

in model M_1 and

$$\hat{p}_i = \frac{x_i}{n}, \quad i = 1, 2$$

in model M_2.

The second derivatives are given by

$$H = \frac{x_1 + x_2}{\hat{p}^2} + \frac{2n - x_1 - x_2}{(1 - \hat{p})^2}$$

in model M_1 and

$$H_i = \frac{x_i}{\hat{p}_i^2} + \frac{n - x_i}{(1 - \hat{p}_i)^2}$$

in model M_2.

Using the above to approximate the Bayes factor first with Laplace approximation

$$K_L = \exp(L(M_1) - L(M_2))$$

and with BIC

$$K_B = \exp(BIC(M_1) - BIC(M_2)).$$

we get $K_L = 1.1585$ and $K_B = 2.4836$ in the data in (1a), and $K_L = 0.0813$ and $K_B = 0.20079$ in (1b). See also `ex2.m` for a Matlab script to calculate these.

Problem 3. *"Model selection for GMM with BIC and cross-validation."*

In many machine learning applications model selection is crucial. In this exercise, you will practice two common approaches for model selection;

1. *Bayesian Information Criterion (BIC), (as an approximation to the 'Bayesian model selection')*
2. *Cross-Validation (as a representative for predictive model selection criteria)*

In both scenarios, you are given a data set which has been simulated using three classes (the true class labels are given for your convenience, but they should not be used in learning the model) and the code for the Gaussian Mixture Model (GMM).

The data has been divided into training and test sets. Use each criterion to select the number of components in the GMM using the training data. Use the selected model to predict the test set. For simplicity, only the number of mixture components of the GMM will be considered when selecting the appropriate model (In practice, other parameters could also be considered). Report also the test set log-likelihood. Furthermore, in your solution, explain the pros and cons of the two approaches and comment which approach you would consider better and why.

Hint: Use `ex7_1_templates.m` and `ex7_2_templates.m` and modify relevant parts into your solution (you don't need to write your own code for fitting the GMM). Also see the function `aicbic` to implement BIC (read `help aicbic` in Matlab) if you like.

Solution. See Matlab code.