Skolkovo Institute of Science and Technology

Practical Assignment. Bayesian Reasoning

Bayesian Methods – Advanced Machine Learning, Spring 2016, Term 3

Start Date: Tuesday, February 2.
Submission Deadline: **Tuesday, February 9, 23:59.**
Maximum points: **5.0** (+ bonus points)
Programming Languages: Python + NumPy.

Questions regarding this assignment should be sent to *bayesml@gmail.com*. Please use the following prefix for the subject: [BMML Skoltech 2016]

Probabilistic Model of Lecture Attendance

Consider a probabilistic model of students attending course lectures. Let the course be a mandatory course for a students, and an optional course for b students. A student attends lecture of a mandatory course with probability p_1, and a lecture of an optional course with probability p_2. Denote by c the number of students who attend the lecture. Then the random variable $c|a, b$ is a sum of two Binomial random variables: $\mathrm{Bin}(a, p_1)$ and $\mathrm{Bin}(b, p_2)$.

Now, suppose the lecturer decides to register attending students. During the lecture, she asks everyone present to write down their names in a list. Each student writes down his own name, and, with probability p_3, additionally writes down his absent friend's name. We assume that no name appears twice in the list. Denote by d the total number of students registered on the lecture. Then the random variable $d|c$ is a sum of c and a Binomial random variable $\mathrm{Bin}(c, p_3)$.

In order to completely define the probabilistic model, we need to specify priors for a and b. We choose discrete uniform priors with support $[a_{min}, a_{max}]$ and $[b_{min}, b_{max}]$, respectively. Thus, we have specified the following probabilistic model:

$$\begin{aligned}
&p(a, b, c, d) = p(d|c)p(c|a, b)p(a)p(b), \\
&d|c \sim c + \mathrm{Bin}(c, p_3), \\
&c|a, b \sim \mathrm{Bin}(a, p_1) + \mathrm{Bin}(b, p_2), \\
&a \sim \mathrm{Unif}[a_{min}, a_{max}], \\
&b \sim \mathrm{Unif}[b_{min}, b_{max}].
\end{aligned} \tag{1}$$

Now, let's simplify the model 1 slightly. We know that when the number of trials n is large, and probability of success p is low, we have, with high accuracy, $\mathrm{Bin}(n, p) \approx \mathrm{Poiss}(\lambda)$, $\lambda = np$. We also know that a sum of two Poisson random variables with parameters λ_1 and λ_2 is a Poisson random variable with parameter $\lambda_1 + \lambda_2$. Thus, we can consider the following approximation of model 1:

$$\begin{aligned}
&p(a, b, c, d) = p(d|c)p(c|a, b)p(a)p(b), \\
&d|c \sim c + \mathrm{Bin}(c, p_3), \\
&c|a, b \sim \mathrm{Poiss}(ap_1 + bp_2), \\
&a \sim \mathrm{Unif}[a_{min}, a_{max}], \\
&b \sim \mathrm{Unif}[b_{min}, b_{max}].
\end{aligned} \tag{2}$$

Assignment

Consider model 2 with parameters $a_{min} = 75$, $a_{max} = 90$, $b_{min} = 500$, $b_{max} = 600$, $p_1 = 0.1$, $p_2 = 0.01$, $p_3 = 0.3$. Perform the following numerical experiments:

1. Find expected value and variance of marginals for all random variables in the model: a, b, c, d.

2. Study how indirect information improves the estimate of c. To do that, plot the distribution and find the expected value and variance for distributions $p(c)$, $p(c|a)$, $p(c|b)$, $p(c|d)$, $p(c|a, b)$, $p(c|a, b, d)$, when the parameters a, b, d equal the expectation of the respective marginals, rounded to the nearest integer.

3. Determine which one of the parameters a, b, d contributes most to improving of the estimate of c (in the sense of the variance of distribution). Check that $\mathbb{D}[c|d] < \mathbb{D}[c|b]$ and $\mathbb{D}[c|d] < \mathbb{D}[c|a]$ for all permissible values of a, b, d. Are the sets $\{\, (a, b) \mid \mathbb{D}[c|b] < \mathbb{D}[c|a] \,\}$ and $\{\, (a, b) \mid \mathbb{D}[c|b] \geq \mathbb{D}[c|a] \,\}$ linearly separable?

4. Measure the time required to estimate the distributions $p(c)$, $p(c|a)$, $p(c|b)$, $p(c|d)$, $p(c|a,b)$, $p(c|a,b,d)$, $p(d)$.

5. Repeat experiments 1-4 for the exact model 1. Compare with results for model 2. Which parameter's estimate, and under what conditions, is most different in models 1 and 2? Explain the result.

Use the following permissible values for random variables: for c $[0, a_{max} + b_{max}]$, for d $[0, 2(a_{max} + b_{max})]$.
The estimation of any single distribution should not take more than 30 seconds.
Bonus: if the estimation takes less than a second, you get 0.5 bonus points.

Submission Guidelines

The assignment is to be submitted via **Canvas**. The submission must contain:

- Report in PDF format or as IPython Notebook with description of the experiments.
- Source code.

The source code should be contained in a module named br_surname. Estimation of different distributions should be implemented as **separate functions**. Function prototype for estimation of distribution $p(c|a,d)$ is presented in table 1. The functions for other distributions should be named in the same fashion. The names of variables after | symbol should be sorted alphabetically.

Late submission policy

The assignment may be submitted late, but with a late submission penalty. The late submission penalty for this assignment is 0.1 points per day, capped at 4 points.

Collaboration

The assignment have to be done individually in the sense that no sharing of code or solutions is allowed. Discussion of the assignment is allowed and encouraged.

Table 1: Python function prototype for estimation of distribution $p(c|a,d)$ for models 1 and 2

p, c = pc_ad(a, d, params, model=2)
INPUT *a* – value of variable *a*; *d* – value of variable *d*; *params* – parameters of probabilistic model, dictionary with keys 'amin', 'amax', 'bmin', 'bmax', 'p1', 'p2', 'p3'; *model* – model number;
OUTPUT *p* – probability distribution, numpy array of length len(c); *c* – distribution support, numpy array.