**ВМК МГУ**

# Семинар 3. Выбор модели по Байесу

**Курс: Байесовские методы в машинном обучении, 2016**

1. Рассмотрим модель BetaMix-Binomial:

$$p(k,q|N,\boldsymbol{w},\boldsymbol{a},\boldsymbol{b}) = p(k|N,q)p(q|\boldsymbol{w},\boldsymbol{a},\boldsymbol{b}),$$
$$p(k|N,q) = C_N^k q^k (1-q)^{N-k},$$
$$p(q|\boldsymbol{w},\boldsymbol{a},\boldsymbol{b}) = \sum_j w_j \,\mathrm{Beta}(q|a_j,b_j),\ a_j,b_j>0,\ w_j\ge 0,\ \sum_j w_j = 1.$$

   Требуется вычислить апостериорное распределение $p(q|k,N,\boldsymbol{w},\boldsymbol{a},\boldsymbol{b})$, а также найти обоснованность модели $p(k|N,\boldsymbol{w},\boldsymbol{a},\boldsymbol{b})$.

2. Для модели BetaMix-Binomial вычислить прогноз для $k_1$ успехов в новых $N_1$ испытаниях, т.е. найти $p(k_1|N_1,k,N,\boldsymbol{w},\boldsymbol{a},\boldsymbol{b})$. Пусть имеется $M$ моделей BetaMix-Binomial, т.е. заданы распределения $p(k,q|N,\boldsymbol{w}_m,\boldsymbol{a}_m,\boldsymbol{b}_m)$, $m=\overline{1,M}$. Требуется вычислить байесовский прогноз $p(k_1|N_1,k,N,\{\boldsymbol{w}_m,\boldsymbol{a}_m,\boldsymbol{b}_m\}_{m=1}^M)$ для всей совокупности моделей.

3. Пусть дискретная случайная величина $x$ принимает значения $1,2,\ldots,l$ с вероятностью $q_1,q_2,\ldots,q_l$ соответственно. Пусть далее в $N$ независимых испытаниях с величиной $x$ значение 1 выпало $k_1$ раз, значение 2 – $k_2$ раз, …, значение $l$ – $k_l$ раз. Требуется найти вероятность данного события $p(k_1,k_2,\ldots,k_l|\boldsymbol{q},N)$, подобратья априорное сопряженное распределение на $\boldsymbol{q}$, найти апостериорное распределение $p(\boldsymbol{q}|\boldsymbol{k},N)$, обоснованность модели $p(\boldsymbol{k}|N)$ и прогнозное распределение $p(\boldsymbol{k}_1|N_1,\boldsymbol{k},N)$.

4. Рассмотрим задачу моделирования уровней смертностей в городах от заданного заболевания. Пусть $N_i$ – население $i$-го города, а $x_i$ – число зафиксированных смертей за определенный период времени, $i=1,\ldots,N$. Пусть $\theta_i$ – уровень смертности в $i$-ом городе. Составим следующую вероятностную модель:

$$p(X,\Theta|\boldsymbol{N},\boldsymbol{\alpha}) = \prod_{i=1}^N p(x_i|\theta_i,N_i)p(\theta_i|\boldsymbol{\alpha}),$$
$$p(x_i|\theta_i,N_i) = C_{N_i}^{x_i}\theta_i^{x_i}(1-\theta_i)^{N_i-x_i},$$
$$p(\theta_i|\boldsymbol{\alpha}) = \mathrm{Beta}(\theta_i|\boldsymbol{\alpha}).$$

   Требуется найти обоснованность модели $p(X|\boldsymbol{N},\boldsymbol{\alpha})$, а также байесовсвкую оценку для параметра $\theta_i$ в виде мат.ожидания $p(\theta_i|X,\boldsymbol{N},\boldsymbol{\alpha})$.

5. Рассмотрим задачу моделирования уровней подготовки в школах по ЕГЭ по заданному предмету Пусть $N_i$ – количество учеников в $i$-ой школе, а $x_{ij}$ – оценка по ЕГЭ $j$-го ученика в $i$-ой школе. Пусть средняя оценка по школе $\theta_i$ и оценка $x_{ij}$ связаны как $p(x_{ij}|\theta_i) = \mathcal{N}(x_{ij}|\theta_i,\beta^{-1})$, где велечина $\beta$ известна. Требуется по аналогии с предыдущей задачей составить вероятностную модель описания данных с введением общего априорного распределения на $\theta_i$ для всех школ, выбираемого в семействе сопряженных. Требуется также записать обоснованность введенной модели и найти байесовскую оценку для $\theta_i$ как мат.ожидания апостериорного распределения $p(\theta_i|X)$.