Задание 3 по курсу «Байесовский выбор моделей»

Общая информация

- Время сдачи задания: 18е ноября, 21:00 по Москве;
- Максимальная базовая оценка за задание 50 баллов, так что при желании можно выполнять не всё;
- Оценка автора наилучшей работы удваивается (с учетом баллов сверх 50), но не более, чем до 125 баллов;
- Вопросы и само задание принимаются по почте: aduenko@phystech.edu;
- Тема письма: вопрос по заданию #3 или решение задания #3;
- Опоздание на неделю снижает оценку в 2 раза, опоздание на час на $0.5^{1/(7\cdot 24)} = 0.41\%$;
- Работы опоздавших не участвуют в конкурсе на лучшую работу;
- Задание не принимается после его разбора и / или после объявления об этом.

Задача 1 (15 баллов). Пусть имеется обучающая и тестовая выборки $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$, $\mathbf{X}_{\text{train}} \in \mathbb{R}^{m_1\times n}$, $\mathbf{y}_{\text{train}} \in [0, 1]^{m_1}$; $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$, $\mathbf{X}_{\text{test}} \in \mathbb{R}^{m_2\times n}$, $\mathbf{y}_{\text{test}} \in [0, 1]^{m_2}$, полученные из общей модели генерации данных с совместным правдоподобием

$$p(\mathbf{y},\, \mathbf{w},\, \mathbf{X}|\mathbf{A}) = \prod_j N(\mathbf{x}_i|\mathbf{0},\, \sigma^2\mathbf{I}_n)N(\mathbf{w}|\mathbf{0},\, \mathbf{A}^{-1})\prod_j p(y_j|\mathbf{x}_j,\, \mathbf{w}),$$

где $p(y_j|\mathbf{x}_j,\, \mathbf{w})$ дается моделью логистической регрессии, то есть

$$\mathbb{P}(y_j = 1) = \frac{1}{1+\exp(-\mathbf{w}^{\mathsf{T}}\mathbf{x}_j)}.$$

а) Выписать формулу для апостериорного распределения $p(\mathbf{w}|\mathbf{X}_{\text{train}},\, \mathbf{y}_{\text{train}},\, \mathbf{A})$ и получить его нормальную аппроксимацию $p(\mathbf{w}|\mathbf{X}_{\text{train}},\, \mathbf{y}_{\text{train}},\, \mathbf{A}) \approx N(\mathbf{w}_0,\, \mathbf{H}_0^{-1})$ (4 балла);
б) Пусть $\hat{\mathbf{p}}$ – вектор оценок вероятностей принадлежности классу 1 для некоторого классификатора на тестовой выборке. Введем уверенность $C(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$ классификатора на тестовной выборке как

$$C(\hat{\mathbf{p}}) = \sum_{i=1}^{m_2} |\hat{p}_i - 0.5|.$$

Рассмотрим так же правдоподобие тестовой выборки относительно вектора $\hat{\mathbf{p}}$ как

$$l(\mathbf{y}_{\text{test}},\, \hat{\mathbf{p}}) = \prod_{i=1}^{m_2} \hat{p}_i^{y^i_{\text{test}}}(1-\hat{p}_i)^{1-y^i_{\text{test}}}.$$

Считая $m_2 = 1000$, а $\sigma^2 = 1$, $\mathbf{A} = \mathbf{I}_n$ известными и фиксированными, для разных размеров обучающей выборки m_1 сравнить с помощью сэмплирования уверенность классификатора на тестовой выборке и правдопобие на ней для точечного MAP-классификатора вида

$$\hat{\mathbf{p}}_{\text{test}} = \frac{1}{1+\exp(-\mathbf{X}_{\text{test}}^{\mathsf{T}}\mathbf{w}_{\text{MAP}})}$$

и для полного байесовского классификатора, учитывающего неопределенность в $\mathbf{w}$ вида

$$\hat{\mathbf{p}}_{\text{test}} = \int \frac{1}{1+\exp(-\mathbf{X}_{\text{test}}^{\intercal}\mathbf{w})} p(\mathbf{w}|\mathbf{X}_{\text{train}},\, \mathbf{y}_{\text{train}}) d\mathbf{w}.$$

Какой практический вывод можно сделать из полученных результатов? (11 баллов)

Задача 2 (20 баллов). Пусть имеется модель линейной регрессии с нормальным шумом

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\, \boldsymbol{\varepsilon} \sim N(0,\, \sigma^2\mathbf{I}),$$

где σ^2 – известно, и априорным распределение на $\mathbf{w}$ $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m},\, \mathrm{diag}(\mathbf{s}))$, где $\mathbf{m}$ и $\mathrm{diag}(\mathbf{s})$ неизвестные гиперпараметры.
а) Выписать совместное правдоподобие $p(\mathbf{y},\, \mathbf{w}|\mathbf{X},\, \mathbf{m},\, \mathbf{s})$, задающее вероятностную модель. (2 балла)
б) Получить апостериорное распределение на вектор $\mathbf{w}$, предполагая $\mathbf{m}$ и $\mathbf{s}$ известными. Что происходит, если $s_i = 0$? (4 баллов)
в) Решить задачу максимизации обоснованности

$$p(\mathbf{y}|\mathbf{X},\, \mathbf{m},\, \mathbf{s}) = \int p(\mathbf{y}|\mathbf{X},\, \mathbf{w}) p(\mathbf{w}|\mathbf{m},\, \mathrm{diag}(\mathbf{s})) d\mathbf{w}$$

по гиперпараметрам $\mathbf{m}$ и $\mathbf{s}$. Какой вывод можно сделать из полученного результата? (14 баллов)

Задача 3 (20 баллов). Пусть имеется две двухсторонние монеты, случайно и независимо выбранные из всех существующих монет достоинством в 2 рубля. Пусть было произведено $n_1 = 10$ бросаний первой монеты и $n_2 = 10000$ бросаний второй. Среди $n_1 = 10$ результатов бросания первой монеты было $k_1 = 3$ орла, а среди $n_2 = 10000$ бросаний второй – $k_2 = 5050$ орлов.
а) Построить вероятностную модель эксперимента, записав правдоподобие и введя априорные распределения на вероятности p_1 и p_2 выпадания орлов для первой и второй монеты соответственно. Опишите, как и из каких соображений Вы выбрали априорные распределения $q(p_1)$ и $q(p_2)$.(4 балла)
б) Получить апостериорные распределения $q(p_1|k_1,\, n_1)$ и $q(p_2|k_2,\, n_2)$. (4 балла)
в) Пусть теперь рассматривается две вероятностные модели: M_1 с $p_1 = p_2 = p$ и априорным распределением, которые было ранее выбрано Вами для p_1 и полная модель M_2 из пункта а), где p_1 и p_2 априорно выбраны независимо из $q(p_1)$ и $q(p_2)$. Сосчитать апостериорную вероятность обеих моделей, считая их априори равновероятными ($p(M_1) = p(M_2) = 0.5$). Какой вывод можно сделать из результата? (12 баллов)

Задача 4 (10 баллов). а) Что такое дивергенция Кульбака-Лейблера (KL-divergence), что она показывает и когда определена? (2 балла)
б) Докажите, что значение дивергенции Кульбака-Лейблера неотрицательно (3 балла).
в) Пусть у Вас есть две модели логистической регрессии с равномерным априорным псевдораспредлением на параметр $\mathbf{w}$, оцененные на двух разных выборках $(\mathbf{X}_1,\, \mathbf{y}_1)$ и $(\mathbf{X}_2,\, \mathbf{y}_2)$ с одинаковым набором из двух признаков. Пусть апостериорные распределение для первой выборки $\mathbf{w} \sim N\left(\mathbf{w} \,|\, [1,\, 1]^{\intercal},\, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$, а для второй выборки $\mathbf{w} \sim N\left(\mathbf{w} \,|\, [-2,\, -3]^{\intercal},\, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix}\right)$. Считая, что выборки сгенерированны с помощью модели логистической регрессии, можно ли с уверенностью утверждать, что истинные векторы параметров этих моделей $\mathbf{w}_1$ и $\mathbf{w}_2$ разные? (5 баллов)