# Задание 3 по курсу «Байесовский выбор моделей»

### Общая информация

- Время сдачи задания: 16е 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 150 баллов;
- Вопросы и само задание принимаются по почте: aduenko1@gmail.com;
- 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

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

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

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

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 $\sigma^2$ – известно, и априорным распределение на $\mathbf{w}$ $p(\mathbf{w}) = N(\mathbf{w}|\mathbf{m},\, \text{diag}(\mathbf{s}))$, где $\mathbf{m}$ и $\text{diag}(\mathbf{s})$ 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

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

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

**Задача 3 (40 баллов).** Рассмотрим задачу двуклассовой классификации $y \in \{\pm 1\}$, где класс -1 означает успешный возврат займа, а 1 – невозврат или возврат с просрочкой.

Пусть по анкете на предоставление займа составляется два признака с действительными значениями $x_1,\, x_2 \in \mathbb{R}$.

Пусть $p(\mathbf{x}|y=-1) = N(\mathbf{0},\, \mathbf{I})$, $p(\mathbf{x}|y=1) = N(\mathbf{e},\, 4\,\mathbf{I})$, где $\mathbf{e} = [1,\, 1]^{\intercal}$, $\mathbf{I} = \text{diag}(\mathbf{e})$.

При заполнении анкеты соискатель либо указывает подлинное значение признака, либо не указывает его вовсе (решение принимается независимо по признакам).

Соискатели из класса -1 указывают значение признака $x_i$ в анкете с вероятностью $p_{-1} = p_0$, не зависящей от значения признака. Соискатели класса 1 указывают значение признака $x_i$ в анкете с вероятностью $p_1 = p_0 \exp(-\lambda \max(0,\, x_i))$, $\lambda \geq 0$.

а) Сгенерировать обучающую и тестовую выборки размера $N_1 = 10000$, $N_2 = 10000$ со сбалансированными классами без пропусков.

б) Для $p_0 = 0.9$ и $\lambda = 0$ сгенерировать бинарную маску пропусков для каждого из классов на обучении и тесте.

в) Обучить логистическую регрессию на обучающей выборке, используя следующие стратегии заполнения пропусков на обучении:

- Среднее этого признака по всем объектам обучения;
- Среднее этого признака по всем объектам обучения того же класса.

г) Сделать прогноз вероятностей принадлежности классу 1 $\mathbf{p}_1^{\text{test}}$ для тестовой выборки с помощью полученной модели, заполняя пропуски средним по всем объектам обучения и контроля.

д) Вычислить AUC, а также логарифм правдоподобия $LL_{\text{test}} = \sum_{i \in X_{\text{test}}} (\log p_i \cdot [y_i = 1] + \log(1-p_i) \cdot [y_i = -1])$ на полной тестовой выборке и на подвыборке, содержащей объекты без пропусков. Осреднить по $Q = 100$ генерациям выборок (4 балла).

е) Повторить шаги б-д для значений $p_0 = 0.1, 0.3, 0.5, 0.8, 0.9, 0.95, 0.99, 1$ и $\lambda = 0, 0.1, 0.25, 0.5, 1, 2$. Как зависит MAP-оценка вектора весов признаков $\mathbf{w}$ логистической регрессии от $p_0$ и $\lambda$? Как зависит от $p_0$ и $\lambda$ неопределенность в $\mathbf{w}$? (8 баллов) Какие проблемы заметны у рассмотренных методов заполнения пропусков и чем они вызваны по Вашему мнению? (3 балла)

ж) Вывести формулу для прогноза $p(y_{\text{test}}|\mathbf{x}_{\text{test}},\ X_{\text{train}},\ y_{\text{train}})$, считая, что $X_{\text{train}} = X_{\text{train}}^{\text{known}} + X_{\text{train}}^{\text{unknown}}$, $\mathbf{x}_{\text{test}} = \mathbf{x}_{\text{test}}^{\text{known}} + \mathbf{x}_{\text{test}}^{\text{unknown}}$ в общем виде (4 балла).
з) Используя формулу, полученную на предыдущем шаге, предложить более корректный метод учета наличия пропусков и сравнить его с предыдущими при тех же $p_0$ и $\lambda$ (10 баллов).
и) Пусть имеется потенциальный заемщик с истинными значениями признаков $x_1$, $x_2$, который знает алгоритм оценки вероятности класса 1, и может принять решение скрыть значения одного или обоих признаков. Предложите схему принятия решения, максимизирующую для него оцененную вероятность принадлежности классу -1 (5 баллов).
й) Предложите схему принятия решения, которая не позволит увеличить оцененную вероятность принадлежности классу -1 путем сокрытия значения признаков? (4 баллов) Какой вывод можно сделать из результатов этого и предыдущего пункта? (2 балла)

**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 псевдораспределением 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 $\mathbf{w} \sim N\left(\mathbf{w} \mid [1,\ 1]^\intercal,\ \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$, а для второй выборки $\mathbf{w} \sim N\left(\mathbf{w} \mid [-2,\ -3]^\intercal,\ \begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}\right)$. 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)