

## Задание 2 по курсу «Байесовский выбор моделей»

### Общая информация

- Время сдачи задания: 22е октября, 23:59 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: aduenko1@gmail.com & iakovlev.kd@phystech.edu (отправлять на обе сразу);
- Тема письма: вопрос по заданию #2 или решение задания #2;
- 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 (10 баллов).** Пусть есть НОР (i.i.d.) выборка  $x_1, \dots, x_n$ ,  $n > 100$  из нормального распределения со средним  $m$  и неизвестной дисперсией  $\sigma$ . На уровне значимости  $\alpha$  проверить гипотезу  $H_0$  о том, что  $m = 0$ . Выписать критическую область и сосчитать мощность критерия  $W$  в зависимости от истинных  $m$  и  $\sigma$ .

**Задача 2 (20 баллов).** Пусть имеется обучающая и тестовая выборки  $(\mathbf{X}_1, \mathbf{y}_1)$ ,  $\mathbf{X}_1 \in \mathbb{R}^{m_1 \times n}$ ,  $\mathbf{y}_1 \in [-1, 1]^{m_1}$ ,  $(\mathbf{X}_2, \mathbf{y}_2)$ ,  $\mathbf{X}_2 \in \mathbb{R}^{m_2 \times n}$ ,  $\mathbf{y}_2 \in [-1, 1]^{m_2}$ , полученные из общей модели генерации данных с совместным правдоподобием

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

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

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

а) Пусть Вам известен настоящий вектор  $\mathbf{w}$ , полученный из априорного распределения  $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1} \mathbf{I}_n)$ . Вычислите ожидаемое максимальное качество в терминах AUC на тестовой выборке при  $m_2 \rightarrow \infty$  сэмплированием (4 балла), аналитически (6 баллов).

б) Пусть Вами случайно выбран некоторый вектор  $\mathbf{w}_0$ , независимо от настоящего  $\mathbf{w}$ . Вычислить в этом случае для разных  $m_2$  ожидаемое качество в терминах AUC  $\mathbb{E}(\text{AUC})$  для разных  $m_2$  сэмплированием (4 балла), аналитически (6 баллов).

**Задача 3 (10 баллов).** В обозначениях задачи 2

а) Доказать, что Ассигасу (ACC) (доля правильно предсказанных классов объектов) частный случай ASY( $\mathbf{P}$ ) (см. определение из практического задания 1) (2 балла);

б) Пусть класс объектов  $y_j$  не зависит от  $\mathbf{x}_j$ , то есть выборка шумовая.

- Построить наилучший прогноз  $\hat{y}_2$  на тестовой выборке в терминах ACC, если  $\mathbb{P}(y_j = 1) = p$ . (2 балла).
- Построить наилучший прогноз  $\hat{y}_2$  на тестовой выборке в терминах ASY( $\mathbf{P}$ ) в общем случае, если  $\mathbb{P}(y_j = 1) = p$ ? (4 балла)

- Как оценить  $p$  по обучающей выборке и что делать, если оценка не отличается значимо от 0.5? (2 балла)

**Задача 4 (25 баллов).** Пусть имеется выборка  $\mathbf{x}_1^0, \dots, \mathbf{x}_{m_1}^0$  объектов класса 0 размера  $m_0$  и выборка  $\mathbf{x}_1^1, \dots, \mathbf{x}_{m_2}^1$  объектов класса 1 размера  $m_1$ . Пусть известно, что признаки независимы в совокупности в обеих выборках, а также, что признаки имеют нормальное распределение с дисперсиями  $\sigma_j^2$ , одинаковой для одного и того же признака в разных классах, и, возможно, разной между признаками. Пусть требуется проверить гипотезу о том, что мат. ожидание значения признака с номером  $j$  совпадает для обоих классов.

а) Пусть  $\sigma_j^2 = \sigma^2$  известно. Проверить гипотезу о равенстве мат. ожиданий на уровне значимости  $\alpha = 0.05$  (3 балла).

б) Та же задача, что и в пункте а, но  $\sigma_j^2 = \sigma^2$  неизвестно (5 баллов).

в) Пусть  $n = 100$ ,  $\sigma_j^2 = j$ . Для каждой пары  $m_1, m_2$  сгенерировать выборку с такими параметрами, сделав мат. ожидания всех признаков кроме  $j^*$  одинаковыми, а для признака  $j^*$  сделать разницу мат. ожиданий равной 1. Считая  $\sigma_j^2$  неизвестными, реализовать метод, предложенный в п. б) и использовать его для проверки гипотез о равенстве мат. ожиданий для каждого из  $n = 100$  признаков (4 балла). Применить поправку на множественное тестирование Бенджамини-Хохберга (2 балла) и изучить зависимость количества ложных положительных и настоящих положительных отклонений гипотезы о равенстве мат. ожидания от  $m_1, m_2$  (6 баллов).

г) Предложите метод решения этой задачи, если признаки не имеют нормального распределения (5 баллов).

**Задача 5 (15 баллов).** Пусть имеется матрица признаков  $\mathbf{X}$  размера  $m \times n$ .

а) Что такое метод главных компонент? Какую задачу он решает? (3 балла)

б) Описать (доказательно) результат применения (какие будут главные компоненты и соответствующие им собственные числа) метода главных компонент к матрице  $\mathbf{X}$ , если  $m > n$ , объекты независимы, а  $\mathbf{x}_j \in \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$  (4 балла).

в) Пусть  $\mathbf{X}$  состоит из  $n - 1$  зашумленной копии некоторого признака  $\chi_1$ , а также из шкалированного признака  $\chi_2$ , то есть  $\mathbf{X} = [\chi_1 + \varepsilon_1, \dots, \chi_1 + \varepsilon_{n-1}, \kappa \chi_2]$ , где  $\chi_1, \chi_2, \varepsilon_1, \dots, \varepsilon_{n-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$  и независимы в совокупности, а  $\kappa > 0$  – коэффициент шкалирования.

Вычислить в зависимости от коэффициента шкалирования  $\kappa$  ожидаемую первую главную компоненту матрицы  $\mathbf{X}$ , а также ожидаемую долю дисперсии, ею объясняемую, аналитически (5 баллов) и сэмплированием (1 балл). Какой практический вывод можно сделать из полученного результата? (2 балла)