Вопросы к экзамену по курсу
«Математические основы теории прогнозирования» 2011

Вопросы по части Ветрова Д.П.

1. Различные постановки задач машинного обучения (классификация, регрессия, кластеризация, идентификация, прогнозирование, поиск закономерностей). Основные проблемы в теории машинного обучения: переобучение, некорректность данных, малый объем обучающей информации.

2. Метод максимального правдоподобия. Его достоинства и недостатки. Нормальное распределение (одномерное и многомерное), его основные свойства.

3. Решение задач условной оптимизации. Правило множителей Лагранжа. Двойственная задача. Выпуклый вариант теоремы Куна-Таккера.

4. Задача восстановления линейной регрессии. Метод наименьших квадратов. Решение несовместных СЛАУ.

5. Задача восстановления линейной регрессии. Вероятностная формулировка метода наименьших квадратов.

6. Логистическая регрессия. Вероятностная постановка. Регуляризация обучения. Итеративный метод наименьших квадратов.

7. Метод опорных векторов. Прямая задача оптимизации, ее свойства.

8. Метод опорных векторов. Двойственная задачи оптимизации. Ядровой переход.

9. Скрытые марковские модели. Обучение СММ с учителем.

10. Алгоритм динамического программирования и его применение в скрытых марковских моделях.

11. EM-алгоритм и его применение для задачи разделения гауссовской смеси.

12. Проблема уменьшения размерности в описании данных. Формулировка метода главных компонент через критерий минимизации ошибки проектирования.

13. Формулировка метода главных компонент через критерий максимизации разброса в данных. Решение задачи идентификации. Выбор размерности редуцированного пространства.

14. Метод главных компонент в ситуации, когда число признаков превышает число объектов. Ядровой метод главных компонент.

Вопросы по части Журавлева Ю.И.

1. В алгоритмах вычисления оценок объекты задаются наборами значений n числовых признаков. В таблице обучения 2 непересекающихся класса $K_1 = \{S_1, \ldots, S_m\}$ и $K_2 = \{S'_1, \ldots, S'_m\}$. Опорными являются все подмножества из k элементов, $k < n$. Функция близости определяется параметрами $\varepsilon_1, \ldots, \varepsilon_n$ и равна 1 тогда и только тогда, когда выполнены все соответствующие неравенства. Веса всех признаков и всех объектов равны 1. При вычислении $\Gamma_j(S)$, $j = 1, 2$ учитывается только близость с коэффициентом 1 к объектам из своего класса. Написать формулы для $\Gamma_1(S)$ и $\Gamma_2(S)$.

2. В алгоритмах вычисления оценок объекты задаются наборами значений n числовых признаков. В таблице обучения 2 непересекающихся класса $K_1 = \{S_1, \ldots, S_m\}$ и $K_2 = \{S'_1, \ldots, S'_l\}$. Характеристические векторы опорных множеств образуют конъюнкцию $x_1 \cdot \cdots \cdot x_r \cdot \bar{x}_{r+1} \cdot \cdots \cdot \bar{x}_t$, $t < n$. Функция близости определяется параметрами $\varepsilon_1, \ldots, \varepsilon_n$ и равна 1 тогда и только тогда, когда выполнены все соответствующие неравенства. Веса всех признаков и всех объектов равны 1. При вычислении $\Gamma_j(S)$, $j = 1, 2$ учитывается с коэффициентом 1 только отсутствие близости к объектам чужого класса. Написать формулы для $\Gamma_1(S)$ и $\Gamma_2(S)$.

3. Даны два объекта $S_1 = (a_1, \ldots, a_n)$ и $S_2 = (b_1, \ldots, b_n)$. Метрики в множествах значений признаков $\rho_1, \ldots, \rho_n$. Параметры, определяющие близость: $\varepsilon_1, \ldots, \varepsilon_n$. Неравенство t: $\rho_t(a_t, b_t) \le \varepsilon_t$. k – целочисленный параметр. Пусть $M = (q_1, \ldots, q_l)$, $l > k$. Функция близости $N(S_1, S_2, M) = 1$ тогда и только тогда, когда из неравенств $\rho_{q_1}(a_{q_1}, b_{q_1}) \le \varepsilon_{q_1}, \ldots, \rho_{q_l}(a_{q_l}, b_{q_l}) \le \varepsilon_{q_l}$ не выполнено не более k. Для скольких множеств M, состоящих не менее, чем из $k+1$ элемента, функция близости равна 1?

4. Дана таблица обучения

$$\left.\begin{array}{l} S_1 = (1,1,0,1,1,0,\ldots,1,1,0,1,1,0) \\ S_2 = (0,0,1,0,0,1,\ldots,0,0,1,0,0,1) \end{array}\right\} K_1,$$
$$\left.\begin{array}{l} S_3 = (0,0,1,1,1,0,\ldots,0,0,1,1,1,0) \\ S_4 = (1,1,0,0,0,1,\ldots,1,1,0,0,0,1) \end{array}\right\} K_2,$$

Распознаваемый объект

$$S = (0,0,1,0,0,1,1,1,0,1,1,0,\ldots,0,0,1,0,0,1,1,1,0,1,1,0).$$

В какой класс будет отнесен объект S тестовым алгоритмом, если все тесты равноценны и совпадение хотя бы с одной строкой тупикового теста дает классу один голос?

5. Дана таблица обучения с $3m$ признаками:

$$\left.\begin{array}{l} S_1 = (1,1,1,0,0,0,1,1,1,0,0,0,\ldots,1,1,1,0,0,0) \\ S_2 = (0,0,0,1,1,1,0,0,0,1,1,1,\ldots,0,0,0,1,1,1) \end{array}\right\} K_1,$$
$$\left.\begin{array}{l} S_3 = (0,0,0,0,0,0,0,0,0,0,0,0,\ldots,0,0,0,0,0,0) \\ S_4 = (1,1,1,1,1,1,1,1,1,1,1,1,\ldots,1,1,1,1,1,1) \end{array}\right\} K_2,$$

В алгоритме вычисления оценок:

(a) $\{\Omega\}_A = \{(1,2,3), (4,5,6), \ldots, (3m-2, 3m-1, 3m)\}$.

(b) Функция близости задается параметрами $\varepsilon_1, \dots, \varepsilon_{2n}$ и равна 1 тогда и только тогда, когда все 3 соответствующих неравенства выполнены.

(c) $w_i = 1, i = 1, \dots, 2n$. Веса объектов $w^1, \dots, w^4$ не заданы.

(d) Значения параметров $x_{00}, x_{11}, x_{01}, x_{10}$ не заданы.

(e) Решающее правило $\Gamma_j(S) > 2, S \in K_j$, $\Gamma_j(S) < 1, S \notin K_j$.

Можно ли подобрать значения параметров $w^1, \dots, w^4, x_{00}, x_{10}, x_{01}, x_{11}$ таким образом, чтобы объект $(0, 0, \dots, 0)$ относился к классу K_2 и не относился к классу K_1, а объект $(1, 1, \dots, 1)$ относился к классу K_1 и не относился к классу K_2?

6. Пусть $S_i = (a_1, \dots, a_n)$ – объект из таблицы обучения, $S = (b_1, \dots, b_n)$ – распознаваемый объект. $\rho_1, \dots, \rho_n$ – метрики для признаков. Функция близости определяется параметрами $\varepsilon_1, \dots, \varepsilon_n, q$ и равна 1 тогда и только тогда, когда из соответствующих неравенств не выполнено не более q. Для скольких подножеств, содержащих признак j, функция близости равна 0, если

   (a) совокупность опорных множеств состоит из всех k элементных подмножеств, $n > k > q$ (сравниваются объекты S_i и S),

   (b) совокупность опорных множеств состоит из всех непустых подмножеств мощности не меньшей $2q + 1$?

7. Пусть таблица обучения и ее информационная матрица имеют вид

$$\begin{array}{l} \\ S_1 = (1, 0, 0, 1, 1, 0, 0, 1, \dots, 1, 0, 0, 1) \\ S_2 = (0, 1, 1, 0, 0, 1, 1, 0, \dots, 0, 1, 1, 0) \\ S_3 = (1, 0, 1, 0, 1, 0, 1, 0, \dots, 1, 0, 1, 0) \\ S_4 = (0, 1, 0, 1, 0, 1, 0, 1, \dots, 0, 1, 0, 1) \end{array} \quad \begin{array}{c} \begin{array}{cc} K_1 & K_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

Контрольная матрица

$$\begin{array}{l} \\ S'_1 = (1, 1, 1, \dots, 1, 1) \\ S'_2 = (0, 0, 0, \dots, 0, 0) \end{array} \quad \begin{array}{c} \begin{array}{cc} K_1 & K_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

Найти алгоритмы вычисления оценок, отмечающие единицы информационной матрицы контроля, и построить корректный для контрольной матрицы алгоритм в алгебраическом замыкании.

8. В материале обучения представлены объекты двух непересекающихся классов K_1, K_2. Объекты – бинарные векторы размерности n. В K_1 включены все объекты, содержащие m единиц и $n - m$ нулей, в K_2 – все объекты, содержащие l единиц и $n - l$ нулей, $m > 1, m > l, m < n$. Рассматриваются алгоритмы вычисления оценок, у которых $\{\Omega\}_A$ состоит из всех k элементных подмножеств множества $\{1, \dots, n\}$, функция близости определяется параметрами $\varepsilon_i = 0, i = 1, \dots, n$ и равна 1 тогда и только тогда, когда выполнены все соответствующие неравенства, $w_i = 1, \forall i$, $x_{11} = 1, x_{00} = x_{01} = x_{10} = 0$, $w^i = 1, S_i \in K_1$. Найти значения весов объектов из класса K_2 так, чтобы объект $(1, 1, \dots, 1)$ получил большую оценку за класс K_2.

9. Объекты из класса K_1 образуют Хеммингов шар радиуса r_1 с центром в точке $(0,\dots,0)$, объекты класса K_2 образуют Хеммингов шар радиуса r_2 с центром в точке $(1,\dots,1)$. Объект S состоит из q единиц и $n-q$ нулей. Рассматриваются все опорные множества из k элементов. Функция близости определяется параметрами $\varepsilon_1=\dots=\varepsilon_n=0$. При каких условиях объект S будет близок к классу K_1 по большему числу подмножеств (число случаев, когда функция близости равна 1, суммируемых по всем опорным подмножествам и всем элементам из обучающей выборки, принадлежащим соответствующему классу)?

10. Пусть заданы параметры $\varepsilon_1,\dots,\varepsilon_n$ и для двух объектов $S=(a_1,\dots,a_n)$, $S'=(b_1,\dots,b_n)$ из n неравенств $\rho_1(a_1,b_1)\le\varepsilon_1,\dots,\rho_n(a_n,b_n)\le\varepsilon_n$ выполнено m неравенств. Рассмотрим все подмножества номеров координат, составленные не менее, чем из k элементов, содержащие координату с фиксированным номером i. Функция близости равна 1, если не выполнено не более t неравенств, $t<k$. Найти число указанных выше подмножеств, для которых функция близости в этом случае равна 1.

11. Дана таблица обучения

$$\left.\begin{array}{l} S_1=(\overbrace{1,1,\dots,1,1}^{2k},\overbrace{0,0,\dots,0}^{2n-2k})\\ S_2=(0,0,\dots,0,0,1,1,\dots,1)\end{array}\right\}K_1,$$
$$\left.\begin{array}{l} S_3=(1,1,\dots,1,1,1,1,\dots,1)\\ S_4=(0,0,\dots,0,0,0,0,\dots,0)\end{array}\right\}K_2,$$

В алгоритме вычисления оценок:

(a) $\{\Omega\}_A$ – все непустые подмножества из 4-х элементов.

(b) Функция близости задается параметрами $\varepsilon_1=\dots=\varepsilon_{2n}=0$ и равна 1 тогда и только тогда, когда все 4 соответствующих неравенства выполнены.

(c) Веса объектов $w^i=1, i=1,\dots,4$. Веса признаков $w_{2k}=1, w_{2k-1}=2, k=1,\dots,n$.

(d) $x_{11}=2, x_{00}=1, x_{01}=x_{10}=-1$.

Найти оценки за классы K_1,K_2 для объектов

$$S_1'=(1,0,1,0,\dots,1,0),$$
$$S_2'=(0,1,0,1,\dots,0,1).$$

12. Пусть таблица обучения с $2k$ признаками и ее информационная матрица имеют вид

$$\begin{array}{ll} S_1= & (\overbrace{1,1,\dots,1}^{k},\overbrace{1,\dots,1}^{k})\\ S_2= & (0,0,\dots,0,0,\dots,0)\\ S_3= & (1,1,\dots,1,0,\dots,0)\\ S_4= & (0,0,\dots,0,1,\dots,1)\end{array}\qquad \begin{array}{cc} K_1 & K_2\\ \end{array}\begin{pmatrix} 1 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\end{pmatrix}$$

Контрольная матрица

$$\begin{array}{l} S_1'=(1,0,1,0,\dots,1,0)\\ S_2'=(0,1,0,1,\dots,0,1)\end{array}\qquad \begin{array}{cc} K_1 & K_2\end{array}\begin{pmatrix} 1 & 0\\ 0 & 1\end{pmatrix}$$

В алгоритме вычисления оценок:

(a) $\{\Omega\}_A$ – все опорные множества из 4-х элементов.

(b) Функция близости определяется параметрами $\varepsilon_1 = \dots = \varepsilon_n = 0$ и равна 1 тогда и только тогда, когда все соответствующие неравенства выполнены.

(c) $w_1 = \dots = w_{2k} = 1$, $x_{11} = 1, x_{10} = x_{01} = x_{00} = 0$.

Найти значения весов объектов, при которых оператор алгоритма отмечает обе единицы в информационной матрице контрольной выборки.

13. Доказать критерий поглощения. Рассмотрим включение

$$\mathcal{N}_K \subseteq \bigcup_{i=1}^{m} \mathcal{N}_{K_i},$$

где K, K_i – элементарные конъюнкции, $\mathcal{N}_K, \mathcal{N}_{K_i}$ – интервалы (множество единиц соответствующей конъюнкции). Пусть дано преобразование $\pi : x_i \to x_j^{\sigma_{ij}}$, $\binom{i}{j}$ – перестановка, $\sigma_{ij} \in \{0, 1\}$. Доказать, что условие выше выполнено тогда и только тогда, когда

$$\mathcal{N}_{\pi(K)} \subseteq \bigcup_{i=1}^{m} \mathcal{N}_{\pi(K_i)}.$$

Используя критерий поглощения, упростить выражение

$$x_1\bar{x}_2 \vee \bar{x}_1 x_2 \vee x_1\bar{x}_3 \vee \bar{x}_1 x_3 \vee x_2\bar{x}_3 \vee \bar{x}_2 x_3.$$

14. Доказать теорему о построении корректного алгоритма, если все единичные координаты контрольной матрицы были отмечены операторами вычисления оценок $B_1, \dots, B_K$. Построить оператор, отмечающий единичные элементы матрицы

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

если операторы B_1, B_2, B_3 построили матрицы оценок

$$\begin{bmatrix} 1 & 0.4 & 0.4 \\ 0.5 & 1.5 & 1 \\ 0.4 & 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1.5 & 1.5 \\ 1.5 & 1.5 & 1 \\ 1.5 & 1 & 0.4 \end{bmatrix}, \quad \begin{bmatrix} 1.5 & 1.5 & 1.5 \\ 1.4 & 0.4 & 1.5 \\ 1.5 & 1.5 & 1.7 \end{bmatrix}.$$