Программа курса: Методы машинного обучения и поиск достоверных закономерностей

1. Область применения методов, основанных на обучении по прецедентам (машинном обучении). Примеры применения. Понятие обучающей выборки. Способ обучения, основанный на минимизации эмпирического риска.
2. Типы задач машинного обучения в зависимости от характера целевой переменной: распознавание, регрессионный анализ, анализ выживаемости
3. Различные метрики для оценивания эффективности в задачах регрессионного анализа и в задачах распознавания.ROC анализ. Основные цели метода. Способ построения ROC кривых
4. Понятие обобщающей способности. Для каких алгоритмов достигается максимум обобщающей способности. Байесовский классификатор.
5. Способы оценки обобщающей способности. Кросс-валидация.
6. Эффект переобучения.
7. Линейная регрессия. Использование метода наименьших квадратов для оценки коэффициентов. Оценка параметров одномерной регрессии.
8. Поиск коэффициентов многомерной регрессии с помощью МНК. Формула для регрессионных коэффициентов. Явление мультиколлинеарности.
9. Свойства оптимальных регрессий.
10. Трёхкомпонентное разложение обобщённой ошибки. Смысл шумовой составляющей, составляющей сдвига и дисперсионной составляющей. Bias-Variance дилемма.
11. Байесовские методы обучения. Аппроксимация с помощью многомерного нормального распределения. Способ обучения.
12. Линейный дискриминант Фишера. Способ обучения.
13. Метод k-ближайших соседей. Способ обучения.
14. Логистическая регрессия. Способ обучения.
15. Распознавания при заданной точности распознавания одного из классов. Оптимальное решение согласно лемме Неймана-Пирсона.
16. Принцип частичной прецедентности. Понятие тупикового теста. Общая схема тестового алгоритма. Обобщение для веществennозначной информации.

17. Понятие тупикового представительного набора. Общая схема алгоритма распознавания, основанного на тупиковых представительных наборах. Обобщение для вещественнозначной информации.
18. Модель Алгоритмов вычисления оценок. Понятия и опорного множества, функции близости, для вычисления оценок за классы. Компактные формулы для оценок в случае, когда признаки равноправны, а мощность опорных множеств фиксирована.Способы обучения для модели АВО.
19. Модель искусственного нейрона. Пецептрон Розенблатта и метод его обучения, условие сходимости.
20. Многослойный перцептрон и его структура. Аппроксимирующая способность многослойных перцептронов. Метод обратного распространения ошибки.
21. Метод опорных векторов. Концепция максимального “зазора”. Сведение к задаче квадратичного программирования. Условия, налагаемые теоремой Каруша-Куна-Таккера. Двойственная задача квадратичного программирования. Опорные вектора и их роль в формирование распознающего алгоритма.
22. Обобщение исходного варианта метода опорных векторов на случай отсутствия линейной разделимости. “Смягчение” условия линейной разделимости с помощью введения дополнительных переменных. Основные отличия от исходного варианта метода.
23. Обобщение метода опорных векторов, позволяющее строить нелинейные разделяющие поверхности.
24. Решающие деревья. Методы обучения. Индексы неоднородности. Критерии остановки ветвления. Методы “подрезки”.
25. Коллективные методы. Обоснование. Ошибка выпуклой комбинации алгоритмов прогнозирования. Простые комитетные методы. Наивный Байесовский классификатор.
26. Коллективные методы основанные на бутстрэп репликациях. Метод ы бэггинг и бустинг.
27. Решающие леса
28. Методы, основанные на голосовании по системам логических закономерностей. Полные и частичные логические закономерности. Методы поиска. Коллективное решение.
29. Метод «Статистически взвешенные синдромы». Оптимальные разбиения в рамках фиксированных моделей. Коллективное решение.

30. Методы кластеризации. Цели кластерного анализа. Метод k-внутригрупповых средних. Иерархические методы кластеризации.
31. Введение в байесовские сети.
32. Методы анализа выживаемости (надёжности). Оценки кривых выживаемости по методу Каплан-Майера. Модель Кокса.
33. Верификация закономерностей. Перестановочные тесты.
34. Проблема множественного тестирования

ПРИМЕРЫ ЗАДАЧИ

Задача 1

Тестирование системы распознавания для распознавания недобросовестных заёмщиков выявило связь между чувствительностью и ложной тревогой, показанную в таблице. Определить, принесёт ли эксплуатация системы к увеличению доходов банка и определить возможный прирост дохода в расчёте на одного заёмщика. Известно, что доход банка на одного заёмщика составляет 40000 денежных единиц, потери в результате отказа заёмщика от платежей составляют 90000 единиц. Доля недобросовестных заёмщиков составляет 20%.

Чувст.	Лож. Тр.
0.02	0.0001
0.12	0.003
0.23	0.1
0.38	0.13
0.47	0.16
0.58	0.22
0.67	0.25
0.78	0.31
0.89	0.43
0.97	0.47
1	0.51

Задача 2

Наивный байесовский классификатор распознает объекты по двум признакам:

бинарный признак $X_1 \in \{a,b\}$

непрерывный признак X_2 , распределённый нормально.

Параметры распределения для классов K_1 и K_2 представлены в таблице:

μ - математическое ожидание для признака X_2;

σ - стандартное отклонение для признака X_2;

$p(a)$, $p(b)$ - вероятности значений a, b, соответственно

P_a - априорная вероятность класса

	μ	σ	$p(a)$	$p(b)$	P_a
K_1	2	1	0.3	0.7	0.7
K_2	-2	2	0.6	0.4	0.3

Выделить на числовой оси области значений показателя X_2 с отнесениям классам K_1 и K_2 при условии, что $X_1 = a$

Задача 3

По данным, представленным в таблице построить кусочно-непрерывную регрессию вида $Y = \alpha_l + \beta_l X \quad при \quad X \le B$ и $Y = \alpha_r + \beta_r X \quad при \quad X > B$ с точкой излома $B = 15$. При поиске регрессионных коэффициентов использовать метод наименьших квадратов с $L2$ регуляризацией по коэффициентам β_l, β_r. Выразить регрессионные коэффициенты, а также среднеквадратичную ошибку модели на обучающей выборке, через множитель C перед функцией штрафа .

Y	1	8	16	15	13	13
X	2	7	13	20	24	29

На экзамене достаточно решить задачи 1, 2
Или задачу 3