

Машинное обучение, осень 2014/2015

Вопросы по лекционной части

1. Метрические алгоритмы классификации. Методы ближайших соседей, парзеновского окна, потенциальных функций.
2. Отбор эталонных объектов в метрических алгоритмах классификации. Алгоритм СТОЛП.
3. Понятия логической закономерности. Критерии информативности: эвристический, статистический и энтропийный, их взаимоотношения, преимущества и недостатки.
4. Разновидности хорошо интерпретируемых логических закономерностей. Бинаризация вещественных признаков, алгоритм выделения информативных зон.
5. Решающий список. Жадный алгоритм синтеза списка.
6. Проблема переобучения в методе ID3, редукция решающих деревьев.
7. Решающее дерево. Метод обучения дерева ID3, его достоинства и недостатки.
8. Линейный классификатор. Понятие отступа объекта. Виды функции потерь. Метод стохастического градиента.
9. Линейный классификатор. Метод стохастического градиента. Правило Хебба. Адаптивный линейный элемент.
10. Метод стохастического градиента, его недостатки и способы их устранения. Способы ускорения сходимости. Регуляризация (редукция весов).
11. Теорема Новикова о сходимости градиентного метода обучения персептрона с правилом Хебба (с доказательством).
12. Метод опорных векторов: принцип оптимальной разделяющей гиперплоскости, формальная постановка задачи, двойственная задача. Понятие опорного вектора.
13. Ядра и спрямляющие пространства в методе опорных векторов. Способы конструктивного построения ядер. Примеры ядер.
14. Многомерная линейная регрессия. Метод наименьших квадратов. Сингулярное разложение. Проблема мультиколлинеарности, способы её устранения.
15. Задача сглаживания. Оценка Надарая-Ватсона. Выбор функции ядра и ширины окна сглаживания. Проблема выбросов, алгоритм LOWESS.

16. Нелинейная параметрическая регрессия. Методы Ньютона-Рафсона и Ньютона-Гаусса.
17. Метод итеративного перевзвешивания для логистической регрессии (IRLS). Сравнение с методом стохастического градиента.
18. Мультиколлинеарность и регуляризация в задачах многомерной линейной регрессии. Гребневая регрессия. Лассо Тибширани.
19. Метод главных компонент. Теорема о минимизации функционала невязки (с доказательством). Связь с сингулярным разложением.
20. Вероятностная постановка задачи классификации. Функционал среднего риска. Ошибки I и II рода. Теорема об оптимальности байесовского решающего правила (с доказательством).
21. Наивный байесовский классификатор. Параметрический (нормальный) и непараметрический (парзеновский) наивный байесовский классификатор.
22. Непараметрическое оценивание плотности по Парзену-Розенблатту. Метод парзеновского окна. Выбор функции ядра. Выбор ширины окна, переменная ширина окна.
23. Нормальный дискриминантный анализ: квадратичный дискриминант. Выборочные оценки параметров многомерного нормального распределения. Виды разделяющей поверхности.
24. Нормальный дискриминантный анализ: линейный дискриминант Фишера. Выборочная оценка ковариационной матрицы.
25. Проблема мультиколлинеарности и регуляризация ковариационной матрицы в линейном дискриминанте Фишера.
26. Задача разделения смеси распределений. EM-алгоритм (без выбора числа компонентов смеси). Формулы для E-шага и M-шага.
27. Выбор начального приближения и числа компонент смеси в EM-алгоритме (можно без вывода формул для M-шага).
28. Смесь многомерных нормальных распределений. Сеть радиальных базисных функций (RBF). Настройка RBF с помощью EM-алгоритма.
29. Логистическая регрессия. Предположение об экспоненциальном распределении. Теорема о линейности байесовского оптимального классификатора (с доказательством).
30. Логарифмическая функция потерь в логистической регрессии (вывод из принципа максимума правдоподобия). Метод стохастического градиента для логистической регрессии, сравнение с правилом Хэбба.
31. Оптимизация порога решающего правила по критерию отношения цены ошибок I и II рода. Кривая ошибок (ROC curve).

32. Методы аппроксимации и регуляризации эмпирического риска в задачах классификации. Их связь с принципом максимума совместного правдоподобия данных и модели. Вероятностный смысл регуляризации. Виды регуляризаторов.
33. Задача кластеризации. Графовые алгоритмы: связные компоненты, кратчайший незамкнутый путь, Форель. Функционалы качества кластеризации.
34. Задача кластеризации. Агломеративные (иерархические) алгоритмы. Формула Ланса-Вильямса и типы расстояний между кластерами. Алгоритм построения дендрограммы.
35. Задача кластеризации. Статистические алгоритмы кластеризации EM и k-means.