

ФИВТ 2017  
Машинное обучение  
Экзаменационные вопросы.

1. Supervised и unsupervised learning. Стандартные задачи (классификация, регрессия, кластеризация). Простые модели (kNN, naïve bayes, linear regression, k-Means). Оценка качества - кросс-валидация, кривые обучения. Переобучение и недообучение, как их детектировать.
2. Метрики качества в задачах классификации и регрессии: accuracy, precision, recall, F1, ROC-AUC, log loss, MSE, MAE, квантильная, MAPE, SMAPE. Когда какая из метрик предпочтительней. К оценке каких параметров распределений ответов приводят различные метрики.
3. Извлечение признаков (на примере текста, изображений, звука) и предобработка признаков (на примере работы с разреженными и категориальными признаками).
4. Решающие деревья. Как работает уже построенное решающее дерево. Рекурсивное построение деревьев. Энтропийный критерий и критерий gini. Классификационные и регрессионные деревья: в чем различия. Настройка гиперпараметров решающего дерева. Преимущества и недостатки деревьев.
5. Общие методы построения композиций: блэндинг, стэкинг, бэггинг и бустинг. Bias-variance trade-off (без вывода). Анализ бустинга и бэггинга с помощью bias-variance trade-off.
6. Бэггинг и Random Forest. Связь корреляции между ответами моделей и качеством модели в бэггинге. Сравнение Random Forest и GBDT (подбор параметров, переобучение).
7. Бустинг и GBM. Выбор параметров в ансамблях решающих деревьев.
8. Линейные модели в задачах классификации и регрессии: функции потерь, регуляризаторы, оптимизационные задачи. Стохастический градиентный спуск.
9. Линейная регрессия. Геометрический и аналитический вывод формулы для весов признаков. Регуляризация в линейной регрессии: гребневая регрессия и LASSO.
10. Логистическая регрессия. Варианты записи оптимизационной задачи. Оценка вероятности принадлежности к классу. Настройка параметров с помощью стохастического градиентного спуска.
11. Метод опорных векторов: оптимизационная задача в условной и безусловной форме. Опорные векторы (достаточно записать и продифференцировать Лагранжиан; оптимизационная задача, выраженная через двойственные переменные - опционально). Идея Kernel Trick.
12. Нейронные сети, обучение (backprop), слои для нейронных сетей (dense, conv, pooling, batchnorm), нелинейности (relu vs sigmoid, softmax), функции потерь (logloss, l2, hinge).
13. Нейронные сети, обучение (backprop), оптимизация для нейронных сетей (sg, msg, nmsg, rmsprop, adam), регуляризация нейросетей (dropout, dropconnect, l1, l2, batchnorm).
14. Нейронные сети, обучение (backprop), современные сверточные слои нейронных сетей (vgg, resnet, inception) и детали обучения (batchnorm, pretrainig).
15. Нейронные сети, обучение (backprop), сверточные слои, перенос стилей (генеративные модели), визуализация, автоэнкодеры.

16. Рекуррентные НС, обучение (backprop tt), отличие от сверточных, разновидности рекуррентных слоев (RNN, LSTM, GRU), аннотация изображений, перевод, диалоговые системы.
17. Задача снижения размерности пространства признаков. Идея метода главных компонент (РСА). Связь РСА и сингулярного разложения матрицы признаков (SVD).
18. Вычисление SVD в пространствах высокой размерности методом стохастического градиента (SG SVD).
19. Идея методов SNE, tSNE, принципиальные отличия от РСА.
20. Задача кластеризации. Аггломеративная и дивизионная кластеризация.
21. Кластеризация с помощью EM-алгоритма (без вывода M-шага). Формула Ланса-Уилльямса.
22. Алгоритмы k-Means и DBSCAN.