

Программа экзамена по машинному обучению

ml-mipt basic, весна 2020

1. Метрики качества классификации: accuracy, balanced accuracy, precision, recall, f1-score (multiclass extensions), ROC-AUC.
2. Метрики качества регрессии: MSE, MAE, R2, другие варианты.
3. Линейная регрессия.
Постановка задачи линейной регрессии. Аналитическое решение МНК, оптимальность оценки, теорема Гаусса-Маркова (формулировка). Градиентное решение задачи линейной регрессии.
4. Логистическая регрессия.
Эквивалентность решений полученных методом максимального правдоподобия и минимизации логистической функции потерь.
5. Bias-Variance tradeoff.
6. Проблема несбалансированных классов.
7. Задача снижения размерности. Алгоритм PCA. Связь с SVD, теорема Эккарта-Янга (формулировка).
8. Понятие информации, информационной энтропии.
Критерии информативности: энтропийный, Джини.
9. Процедура bootstrap. Бэггинг. Метод случайных подпространств. Смесь моделей и смесь экспертов (декларативно).
10. Random Forest.
11. Бустинг. Градиентный бустинг.
12. Матричные вычисления. Матричное дифференцирование. Производные основных функций: $a^T x$, Ax .
13. Backpropagation.
Градиентный спуск (GD). Стохастический градиентный спуск (SGD).
Adaptive gradient methods. Adagrad, adamax, adadelta. RMSprop. Adam.
14. Neural network concept. Fully-connected networks.
Logistic regression as simple NN.
XOR problem.
15. Losses for NNs: logistic loss, cross-entropy. Softmax and Log-softmax activations, numerical stability.
16. Activation functions, their impact on the network, computational complexity.
17. Matrix convolution. Convolutional layer, backpropagation through it. 1x1 convolutions, comparison to Dense layers. Max/Average Pooling.
18. Seq2 something. Recurrence as proxy to work with local context.
Backpropagation through RNN.
19. LSTM/GRU, memory concept, gates ideas.
20. Vanishing gradient problem. Potential solutions.

21. Работа с категориальными признаками и пропущенными значениями.
Mean encoding.
22. Наивный байесовский классификатор.
23. Подбор гиперпараметров моделей. Кросс-валидация. Утечки в процессе обучения.
24. Проблема переобучения, способы борьбы с ней.
25. Регуляризация в Supervised learning.
L1 и L2 регуляризация, их вероятностная интерпретация. Другие способы регуляризации.
Регуляризация как ограничения/prior на модель (e.g. глубина и кол-во деревьев, Dropout, Batch-normalization, Weights normalization, Data augmentation etc.)

Данные вопросы не попали в основную лекционную программу не будут оцениваться на экзамене, но на них их полезно уметь отвечать.

- *Text mining: Bag of Words, TF-IDF.*
- *Word2vec. Skip-gram, negative sampling, treating idioms as “words”.
Word2vec as matrix factorization (optional).*
- *Геометрические методы машинного обучения и кластеризация. IsoMap, LLE, DBSCAN, k-means, t-SNE*

Теоретический минимум

1. *Постановка задачи обучения с учителем (supervised learning). Отличие регрессии от классификации.*
2. *Что такое объект, целевая переменная, признак, модель, функционал ошибки и обучение?*
3. *Что означает свойство i.i.d.?*
4. *Запишите формулы для линейной модели регрессии и для среднеквадратичной ошибки.*
5. *Что такое градиент? Какое его свойство используется при минимизации функций?*
6. *Запишите формулу для одного шага градиентного спуска. Как модифицировать градиентный спуск для очень большой выборки?*
7. *Что такое кросс-валидация? На что влияет количество блоков в кросс-валидации?*
8. *Чем гиперпараметры отличаются от параметров? Что является параметрами и гиперпараметрами в линейных моделях и в решающих деревьях?*
9. *Что такое регуляризация? Чем на практике отличается L1-регуляризация от L2?*
10. *Запишите формулу для линейной модели классификации. Что такое отступ?*

11. Что такое точность и полнота?
12. Что такое ROC-AUC? Как построить ROC-кривую?
13. Запишите функционал логистической регрессии. Как он связан с методом максимума правдоподобия?
14. Опишите жадный алгоритм обучения решающего дерева.
15. Почему с помощью решающего дерева можно достичь нулевой ошибки на обучающей выборке без повторяющихся объектов?
16. Что такое *bagging*?
17. Что такое случайный лес? Чем он отличается от бэггинга над решающими деревьями?
18. Как в градиентном бустинге обучаются базовые алгоритмы?
19. Зачем нужен *backprop*, что такое производная вектора по вектору?
20. Опишите принцип работы сверточного слоя (CNN).
21. Опишите принцип работы базового рекуррентного слоя (RNN).
22. Что такое *dropout*?
23. Как *Dropout* и *Batch Normalization* меняют свое поведение при эксплуатации модели (в режиме *inference*)?
24. Теорема Байеса.