Программа экзамена по курсу “Машинное Обучение” (АТП ФИВТ)

Компоненты оценки на экзамене (по 15-ти балльной шкале):

- ответ на билет - 1.5 балл;
- ответ на теорвопрос - 0.5 балла;
- ответ на допвопрос - 1 балл.

Регламент проведения экзамена

1. выдается 1 билет; время на подготовку 40 минут
2. во время подготовки к студенту подходит преподаватель и задаёт вопрос и теоретического минимума;
    a. неправильный ответ на вопросы теоретического минимума - автоматический “неуд”.
3. по истечении 40 минут ответ по билету: время на ответ 10 минут
4. после ответа на билет задаётся доп вопрос;
5. суммарная оценка за экзамен суммируется с результатами за семестр, сумма умножается на 2/3, и после округления[1] выставляется оценка за курс.

Вопросы из теоретического минимума не требуют времени на размышление и спрашиваются до получения билета (могут также спрашиваться во время ответа по билету).

[1] По умолчанию математическое округление, но в индивидуальном порядке правило может быть изменено на усмотрение семинариста.

Билеты

1. Основы машинного обучения

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. Извлечение признаков (на примере текста, изображений, звука) и предобработка признаков (на примере работы с разреженными и категориальными признаками).

2. Метрики качества в задачах классификации и регрессии: accuracy, precision, recall, F1, ROC-AUC, log loss, MSE, MAE, MAPE. Когда какая из метрик предпочтительней? К оценке каких параметров распределений ответов приводят MSE и MAE (с обоснованием).

2. Деревья решений

3. 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.

4. 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.

5. Бэггинг и Random Forest. Связь корреляция между ответами моделей и качеством модели в бэггинге.

6. Бустинг и GBM. Выбор параметров в ансамблях решающих деревьев. Сравнение Random Forest и GBDT (подбор параметров, переобучение). 3. Линейные модели

3. Линейный методы

7. Линейные модели в задачах классификации и регрессии: функции потерь, регуляризаторы, оптимизационные задачи. Стохастический градиентный спуск.

8. Линейная регрессия. Геометрический и аналитический вывод формулы для весов признаков. Регуляризация в линейной регрессии: гребневая регрессия и LASSO.

9. Логистическая регрессия. Варианты записи оптимизационной задачи. Оценка вероятности принадлежности к классу. Настройка параметров с помощью стохастического градиентного спуска.

10. 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.

4. Нейронные сети

11. Нейронные сети, обучение (backprop), слои для нейронных сетей (dense, conv, pooling, batchnorm), нелинейности (relu vs sigmoid, softmax), функции потерь (logloss, l2, hinge)

12. Нейронные сети, обучение (backprop), оптимизация для нейронных сетей (sg, msg, nmsg, rmsprop, adam), регуляризация нейросетей (dropout, dropconnetc, l1, l2, batchnorm)

13. Нейронные сети, обучение (backprop), современные архитектуры сверточных нейронных сетей (vgg, resnet, inception) и детали обучения (batchnorm, pretraining), автоэнкодеры

14. Рекуррентные нейросети, обучение (backprop tt), отличие от сверточных, разновидности рекуррентных слоев (RNN, LSTM, GRU), аннотация изображений, перевод, диалоговые системы

5. Кластеризация и уменьшение размерности

15. Задача снижения размерности пространства признаков. Идея метода главных компонент (PCA). Связь PCA и сингулярного разложения матрицы признаков (SVD). Идея методов SNE, tSNE, принципиальные отличия от PCA.

16. Задача кластеризации. Агломеративная и дивизионная кластеризация. Формула Ланса-Уилльямса.

17. Кластеризация с помощью EM-алгоритма (без вывода M-шага). Алгоритмы k-Means и DBSCAN.

Теоретический минимум

1. В чем разница между задачами классификации, кластеризации, регрессии, уменьшения размерности, приведите примеры.

2. Что такое объект, целевая переменная, признак, модель, функционал ошибки и обучение?

3. Запишите формулы для линейной модели регрессии и для среднеквадратичной ошибки.

4. Что такое градиент? Какое его свойство используется при минимизации функций?

5. Запишите формулу для одного шага градиентного спуска. Как модифицировать градиентный спуск для очень большой выборки? Почему в задаче классификации получается получить несмещенную оценку на градиент? Как выглядит эта оценка?

6. Что такое кросс-валидация? На что влияет количество блоков в кросс-валидации?

7. Чем гиперпараметры отличаются от параметров? Что является параметрами и гиперпараметрами в линейных моделях и в решающих деревьях?

8. Что такое регуляризация? Чем на практике отличается L1-регуляризация от L2?

9. Запишите формулу для линейной модели классификации. Что такое отступ?

10. Что такое точность и полнота?

11. Что такое ROC-AUC? Как построить ROC-кривую?

12. Запишите функционал, оптимизируемый в логистической регрессии. Как он связан с методом максимума правдоподобия? (объяснение через экспоненциальные семейства допустимо, но не требуется)

13. Запишите задачу метода опорных векторов для линейно неразделимого случая. Как функционал этой задачи связан с отступом классификатора?

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 бэггинг?

17. Что такое случайный лес? Чем он отличается от бэггинга над обычными решающими деревьями?

18. Как в градиентном бустинге обучаются базовые алгоритмы? Что такое сокращение шага?

19. Зачем нужен backprop, что такое производная вектора по вектору?

20. Чем нейросеть отличается от линейной модели? Приведите примеры нейросетей.

21. Объясните идею weight sharing на примере сверточного слоя. Почему эта техника работает именно с изображениями? Какое свойство входных объектов мы учитываем?

22. В чем отличие между сверточными и рекуррентными слоями?

23. Как работает метод K-Means?

24. Как работает метод t-SNE?

25. Запишите постановку задачи в методе главных компонент.

Вопросы для самоконтроля

1. Основы машинного обучения Основные понятия:

1. Что такое задачи классификации, кластеризации и регрессии? Какие из них относятся к supervised learning, а какие - к unsupervised?

2. Что такое переобучение и недообучение? Как их можно детектировать?

3. Что такое обучающая и тестовая выборки, кросс-валидация? Как устроена k-fold cross validation? Простые методы:

4. Как работает kNN в задаче классификации?

5. Как работает kNN с весами объектов в задаче классификации и в задаче регрессии?

6. Как работает наивный байесовский классификатор, в чем заключается его наивность?

7. Как приближается исходная зависимость y от x в линейной регрессии и как настраиваются веса в ней? Метрики качества в задачах классификации и регрессии:

8. Как вычисляются и в каких задачах (классификации/регрессии) применяются метрики: accuracy, precision, recall, F1-measure, ROC-AUC, log loss, MSE, MAE, RMSE?

9. Решается задача бинарной классификации (с двумя классами 0 и 1), в которой при- меры из класса 0 составляют 95% выборки. Какие метрики из перечисленных в предыдущем вопросе предпочтительней использовать?

10. К оценке какой величины для распределения y при условии x приводят MSE и MAE?

11. Можно ли при таргетах из множества Y = {0; 1} использовать для оценки P (y = 1|x) не log loss, а MSE?

2. Деревья решений

13. Как выглядит решающее дерево?

14. Как применяется уже построенное для задачи классификации дерево? А для задачи регрессии?

15. Как строятся решающие деревья? (рекомендуется обратиться к материалам лекций или документации sklearn)

16. Как выглядят энтропийный критерий, критерий Джини и среднеквадратичное отклонение, используемое как критерий в задаче регрессии?

17. Что такое node impurity и goodness of split? Как они связаны?

18. Какие преимущества и недостатки есть у деревьев? (полезно как подумать самостоятельно, так и обратиться к документации sklearn)

19. Есть ли разница (с точки зрения вида получаемого в итоге дерева): строить каждое разбиение в дереве, максимизируя информативность, или строить каждое разбиение, минимизируя <<ошибку>>, как было предложено на первой лекции про деревья? Общие идеи построения композиций

20. Что такое bagging, blending, stacking, boosting?

21. Нужно ли как-то делить выборку, чтобы избежать переобучения, при реализации стэкинга?

22. В чем преимущества и недостатки бустинга и бэггинга? Градиентный бустинг

23. В чем основная идея градиентного бустинга?

24. Как выглядит алгоритм градиентного бустинга в самом общем виде --- с произвольной функцией потерь в функционале ошибки и произвольным функционалом, оценивающим качество приближения антиградиента?
25. Как выглядит алгоритм градиентного бустинга с квадратичными функциями потерь? На что настраиваются базовые алгоритмы?
26. Как выглядит алгоритм градиентного бустинга в случае задачи бинарной классификации?
27. Какие параметры есть у классификаторов и регрессоров на основе градиентного бустинга над деревьями в sklearn и XGBoost? Какие параметры стоит настраивать в первую очередь?
28. Какая высота деревьев оправдана в градиентном бустинге над деревьями? Почему?
29. Есть ли у градиентного бустинга склонность к сильному переобучению при увеличении количества деревьев? С какими еще параметрами алгоритма эффект переобучения может быть связан?
30. Какие есть методы борьбы с переобучением, применяемые в градиентном бустинге? Случайный лес
31. Как работает Random Forest?
32. Зачем в Random Forest делается рандомизация с выбором подмножества признаков в каждом сплите?
33. Какой высоты деревья стоит строить в Random Forest?
34. Какие параметры есть у Random Forest в sklearn? Какие параметры стоит настраивать в первую очередь?
35. Есть ли у Random Forest склонность к сильному переобучению при увеличении количества деревьев?
36. С какими еще параметрами алгоритма эффект переобучения может быть связан?
37. Как работает ExtraTreesClassifier из sklearn? 3. Линейные модели

3. Линейные методы

38. Как выглядит решающее правило в линейной классификации? А зависимость, которой мы приближаем ответы в линейной регрессии?
39. Что такое функции потерь в задачах классификации и регрессии? Зачем они нужны?
40. Что такое регуляризаторы? Какими они бывают в задачах линейной классификации и регрессии? Зачем они нужны?
41. Как в общем виде выглядит оптимизационная задача в линейной классификации или линейной регрессии?
42. Как работает настройка весов в линейной модели с помощью SGD (Stochastic Gradient Descent)? Как выглядит правило обновления весов?
43. Учитывается ли коэффициент сдвига w_0 в регуляризаторе? Почему?
44. Почему линейные модели рекомендуется применять к выборке с нормированными значениями признаков?
45. Как выглядит оптимизационная задача в логистической регрессии? А в SVM?

46. Выпишите и докажите формулу для весов в линейной регрессии (с квадратичной функцией потерь). То же самое для гребневой регрессии.
47. Выпишите SGD для логистической регрессии с ℓ_2-регуляризацией и для SVM с линейным ядром.
48. В чем заключается идея ядер в SVM?
49. Какие преимущества и недостатки есть у линейных моделей?

4. Нейронные сети

50. Чем нейросети отличаются от линейных моделей а чем похожи?
51. В чем недостатки полносвязных нейронных сетей какая мотивация к использованию свёрточных?
52. Какие слои используются в современных нейронных сетях? Опишите как работает каждый слой и свою интуицию зачем он нужен.
53. Может ли нейросеть решать задачу регрессии, какой компонент для этого нужно заменить в нейросети из лекции 1?
54. Почему обычные методы оптимизации плохо работают с нейросетями? А какие работают хорошо? Почему они работают хорошо?
55. Для чего нужен backprop, чем это лучше/хуже, чем считать градиенты без него? Почему backprop эффективно считается на GPU?
56. Почему для нейросетей не используют кросс-валидацию, что вместо неё? Можно ли ее использовать?
57. Чем отличаются современные сверточные сети от сетей 5-летней давности?
58. Какие неприятности могут возникнуть во время обучения современных нейросетей?
59. У вас есть очень маленький датасет из 100 картинок, классификация, но вы очень хотите использовать нейросеть, какие неприятности вас ждут и как их решить? что делать если первый вариант решения не заработает?
60. Как сделать стайл трансфер для музыки? оО
61. Можно ли использовать сверточные сети для классификации текстов? Если нет обоснуйте, если да, то как? как решить проблему с произвольной длиной входа?
62. Чем LSTM лучше/хуже чем обычная RNN?
63. Выпишите производную $\frac{d c_{n+1}}{d c_{k}}$ для LSTM http://colah.github.io/posts/2015-08-Understanding-LSTMs/, объясните формулу, когда производная затухает, когда взрывается?
64. Зачем нужен TBPTT почему BPTT плох?
65. Как комбинировать рекуррентные и сверточные сети, а главное зачем? Приведите несколько примеров реальных задач.
66. Объясните интуицию выбора размера эмбеддинг-слоя? почему это опасное место?

5. Кластеризация и уменьшение размерности

68. В чём заключается проблема мультиколлинеарности?

69. Какие проблемы при обучении алгоритмов возникают из-за большой размерности пространства признаков?
70. В чем суть проклятия размерности?
71. Какая связь между решением задачи PCA и SVD-разложением матрицы регрессии?
72. Почему в tSNE расстояние между парами объектов измеряется "по Стьюденту" и как это помогает решить проблему "скрученности" (crowding problem)?
73. На какой идее базируются алгоритмы аггломеративной кластеризации? Напишите формулу Ланса-Вильма.
74. Какие два шага выделяют в алгоритме кластеризации k-means?
75. В чём отличия (основные упрощения) k-means от EM-алгоритма кластеризации?
76. Какой принцип работы графовых алгоритмов кластеризации? 1
77. В чем некорректность постановки задачи кластеризации? 1
78. Какие 3 интерпретации оптимизационной задачи можно предложить в методе PCA?