Задание 4. Теория

Александр Захаров, 494

11 апреля 2017 г.

1 3.1 Знакомство с линейным классификатором

1.1 Как выглядит бинарный линейный классификатор?

Бинарный линейный классфикатор в классы ± 1: $sign((w, x_i))$. Где w - веса признаков

1.2 Что такое отступ алгоритма на объекте? Какие выводы можно сделать из знака отступа?

Отступ $y_i(w, x_i)$. Показывает, то насколько далек объект от разделяющей плоскости. Показывает, "уверенность"алгоритма в правильности классификации этого объекта.

1.3 Как классификаторы вида $a(x) = sign(< w, x > -w_0)$ сводят к классификаторам вида $a(x) = sign(< w, x >)$?

$< w, x_i > -w_0$ переходит в $< w, x_i >$. Добавлением -1 на 0 место в x_i, а w_0 на 0 место в w

1.4 Как выглядит запись функционала эмпирического риска через отступы? Какое значение он должен принимать для «наилучшего» алгоритма классификации?

Эмпирический риск - число объектов на которых мы ошиблись. $\sum\limits_{i=1}^{n} = I(M_i < 0)$. У идеального классификатора $= 0$

1.5 Если в функционале эмпирического риска (риск с пороговой функцией потерь) всюду написаны строгие неравенства $(M_i < 0)$ можете ли вы сразу придумать параметр w для алгоритма классификации $a(x) = sign(< w, x >)$, минимизирующий такой функционал?

Минимизируем риск занулив все коэффициенты

1.6 Запишите функционал аппроксимированного эмпирического риска, если выбрана функция потерь $L(M)$

$\sum\limits_{i=1}^{n} L(M_i)$

1.7 Что такое функция потерь, зачем она нужна? Как обычно выглядит ее график?

Функция потерь нужна, чтобы мы могли понять насколько ошибается наш классификатор. Обычно $L(x) > 0$ при $x < 0$ и близка к 0 при $x > 0$.

1.8 Приведите пример негладкой функции потерь.

Кусочно линейная функция потерь не гладкая(не диффиренцируема в точке 1) L(x) = max((1 - x), 0)

1.9 Что такое регуляризация? Какие регуляризаторы вы знаете?

Регуляризая - это наложение штрафа на величину коэффициентов модели. l1 (сумма модулей коэффициентов) и l2 (сумма квадратов коэффициентов).

1.10 Как связаны переобучение и обобщающая способность алгоритма? Как влияет регуляризация на обобщающую способность?

При переобучение обобщающая способность алгоритма мала. Он хорошо отвечает только на объектах из обучающей выборки. Регуляризация помогает бороться с этим, то есть повышает обобщающую способность

1.11 Как связаны острые минимумы функционала аппроксимированного эмпирического риска с проблемой переобучения?

Острые минимумы функции эмпирического риска соответствуют тому, что даже при малом изменении параметров модели качество падает значительно. Скорее всего это связано с тем, что именно при таких параметрах мы зануляем функцию потерь для многих объектов из обучающей выборки. При этом слабые изменения параметров значительно ухудшают модель. Значит, такое решение очень сильно подстроилось под обучающую выборку. Значит, модель переобучилась

1.12 Что делает регуляризация с аппроксимированным риском как функцией параметров алгоритма?

Функция риска записывается, как $\sum_{i=1}^{n} L(M_i) + \gamma R(w)$

1.13 Для какого алгоритма классификации функционал аппроксимированного риска будет принимать большее значение на обучающей выборке: для построенного с регуляризацией или без нее? Почему?

На обучающей выборке функционал риска будет больше у модели с регуляризатором, так как она меньше подстраивается под обучающую выборку и пытается выявить какое-то обобщение. А переобученная модель без регуляризации может смочь очень качественно подстроиться под известные объекты

1.14 Для какого алгоритма классификации функционал риска будет принимать большее значение на тестовой выборке: для построенного с оправдывающей себя регуляризацией или вообще без нее? Почему?

Аналогично, из-за большей обобщающей способности можели с регуляризатоом функционал риска для нее будет меньше на тестовой выборке.

1.15 Что представляют собой метрики качества Accuracy, Precision и Recall?

- **Accuracy** - доля правильно классифицированных объектов
- **Precision** - отношение правильно классфицированных объектов классов 1 к количеству объектов классифицированных, как класс 1.
- **Recall** - отношение правильно классфицированных объектов классов 1 к количеству объектов класса 1.

1.16 Что такое метрика качества AUC и ROC-кривая?

Метрика ROC-AUC площадь под ROC-кривой.

ROC-кривая строится по точкам P(p). p -вероятность взятая в качестве граничной, если ответы классификатора рассматривать как вероятность принадлежности к классу 1. P(p) - (FPR, TPR), где FPR - доля объектов класса 0, отнесенных к классу 1, а TPR - доля объектов класса 1, отнесенных к классу 1. Изменяя p от 1 до 0, соединяем P(p) линией, это и есть ROC-кривая.

1.17 Как построить ROC-кривую (нужен алгоритм), если например, у вас есть правильные ответы к домашнему заданию про фамилии и ваши прогнозы?

```
probabilities = algo.predict(x)
AUC = 0.0
FPR = 0.0
TPR = 0.0
zero_class = sum(I[y_i = 0])
first_class = sum(I[y_i = 1])
for predicted, class in sorted(zip(probabilities, answer), key=lambda x: -x[0]):
        if class == 0:
                FPR += 1 / zero_class
                AUC += 1 / zero_class * TPR
        else:
                TPR += 1 / first_class
print(AUC)
```

2 3.2 Вероятностный смысл регуляризаторов

Покажем, что регуляризация задает априорное распределение для параметров модели.

Запишем функцию риска $\sum_{i=1}^{n} L(M_i)+\gamma R(w) \to min$ Аналогичная оптимизация $\sum_{i=1}^{n} -L(M_i)-\gamma R(w) \to max$

$$\sum_{i=1}^{n} -L(M_i) - \gamma R(w) = log\left(e^{\left(\sum_{i=1}^{n} -L(M_i)\right)}\right) + log\left(e^{-\gamma R(w)}\right) = log\left(\prod_{i=1}^{n} e^{-L(M_i)}e^{-\gamma R(w)}\right)$$

Так как логарифм монотонная функция, то предыдущая задача эквивалента следующей оптимизационной задаче

$\prod_{i=1}^{n} e^{-L(M_i)}e^{-\gamma R(w)} \to max$ Ну а это можно рассматривать как максимизацию правдоподобия с точностью до константы, в которой регуляризация является априорной вероятностью.

- l_1: $e^{-\gamma \sum abs(w_i)}$ - многомерное распределение Лапласа с независимыми координатами.
- l_2: $e^{-\gamma \sum w_i^2}$ - многомерное нормальное распределение с независимыми координатами.

Плотности обоих распредение даны с точностью до константы.

3 3.3 SVM и максимизация разделяющей полосы

Предположим что выборка линейно разделима. Тогда $\exists w, b : \forall i(< w, x_i > +b) * y_i > 0$. Так как объектов конечное число, и мы можем варьировать w домножая его на константу, то можно написать $\exists w, b \forall i : (< w, x_i > +b) * y_i > 1$. Ширина полосы при этом $\frac{2}{\|w\|}$. Таким образом переходим к оптимизационной задаче $\|w\| -> min$ при условии $\forall i : (< w, x_i > +b) * y_i > 1$.

В случае линейно неразделимой выборки будем штрафовать, за заход разделяющую полосу таким образом мы получаем оканчательную оптимизационную задачу

$0.5\|w\| + \sum_{i=1}^{n} F_i -> min$ при условии $\forall i : (< w, x_i > +b) * y_i > 1 - F_i$ и $\forall i : F_i \geq 0$.

4 3.4 Kernel trick

Нам необходимо, чтобы $K(x,w)+b = x_1^2 + 2x_2^2 - 3$. Таким образом нам хватит пространства размерностью 2: $\psi(x_1,x_2) = (x_1^2, x_2^2)$. $b = -3$. $w = (1, \sqrt{2})$

5 3.5 l_1 -регуляризация

Запишем оптимизационную задачу как $f(x,w) -> min$ при условии $\|w\|_1 < a$ или же $\|w\|_1 - a < 0$. Так как 1-норма выпукла то это выпуклая задача. При a > 0 Она удовлетворяет условию Слейтера, так как существует вектор с 1-нормой меньше некоторой константы a. Безусловная задача похожа на задачу с регуляризатором $L(w,\lambda) = f(x,w) + \lambda * (\|w\|_1 - a) \to min$. Так как регуляризация существенна, то в $\overline{w} = argmin(L(w,\lambda))$ $\|w\|_1 = a$, а значит $\lambda(\|w\|_1 - a) = 0$. В допустимой области $(\|w\|_1 - a) \leq 0$. Значит по теореме Куна-Таккера решения условной и безусловной задач совпадают. То есть при правильно подобранной константе регуляризации l_1 ограничение нормы и добавление l_1 регуляризатора приводит к одинаковым решениям.

6 3.6 Повторение: метрики качества

Ответы даны в секции 3.1