Машинное обучение. Теоретические задания.

Андрей Карямин, 496 грууппа

Апрель 2017

1 Знакомство с линейным классификатором.

1. Как выглядит бинарный линейный классификатор? (Формула для отображения из множетсва объектов в множество классов.)

$$a(x) = \begin{cases} 0, & f(x) > 0 \\ 1, & f(x) < 0 \end{cases} \tag{1}$$

2. Что такое отступ алгоритма на объекте? Какие выводу можно сделать из знака отступа?

$$M_i = y_i f(x_i) \tag{2}$$

Отрицательный отступ означает то, что классификатор ошибся.

3. Как классификаторы вида $a(x) = sign(<\omega, x> -\omega_0)$ сводят к классификаторам вида $a(x) = sign(<\omega, x>)$?

К вектору x они добавляют фиктивный признак $x_0 = 1$, К весам добавляют ω_0

$$\omega_0 + <\omega, x> \leftrightarrow <\omega', x'> \tag{3}$$

4. Как выглядит запись функционала эмпирического риска через отступы? Какое значение он должен принимать для "наилучшего"алгоритма классификации?

$$Q = \frac{1}{l}\sum_{i=1}^{l} I[a(x_i) \neq y_i] = \frac{1}{l}\sum_{i=1}^{l} I[M_i \leq 0] \tag{4}$$

Для самого топового алгоритма Q = 0, т.к. он не ошибётся.

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

Если положить $\omega = 0$, то $M_i = 0$ и $Q = 0$;

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

$$Q = \frac{1}{l}\sum_{i=1}^{l} L(M) \tag{5}$$

7. Что такое функция потерь, зачем она нужна? Как обычно выглядит её график?

L(a,x) функция потерь - это то насколько алгоритм a ошибается на объекте x.

8. Приведите пример негладкой функции потерь.

$$L(M) = [1 - M]_+ \tag{6}$$

9. Что такое регуляризация? Какие регуляризаторы вы знаете?

   Регуляризатор штрафует за большие веса признаков. Штраф за сложность:
   l_1 - регуляризатор :
   $$\tau \cdot \sum_{k=1}^{n} |\omega_k| \tag{7}$$
   l_2 - регуляризатор :
   $$\tau \cdot \sum_{k=1}^{n} \omega_k^2 \tag{8}$$
   τ - это параметр регрессии.

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

    $Q(a(x^l), x^k)$ - обобщающая способность алгоритма. x^l, x^k - непересекающиеся выборки. Переобучение влечет его большое значение. Регуляризация в свою очередь делает так, чтобы веса признаков не выходили за определённые рамки (при выходе из которых алгоритм переобучается)

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

    В точке острого минимума функционала аппроксимированного риска получается сильный прирост значения этого функционала. Поскольку этот параметр затмевает все остальные, получается переобучение.

12. Что делает регуляризация с аппроксимированным риском как функцией параметров алгоритма?

    Аппроксимированный риск увеличивается при приближении или выходе параметров за определённые границы.

13. Для какого алгоритма классификации функционал аппроксимированного риска будет принимать большее значение на обучающей выборке: для простроенного с регуляризацией или без неё? Почему?

    Поскольку регуляризация увеличивает значение функционала, то с регуляризацией.

14. Для какого алгоритма классификации функционал риска будет приниимать большее значение на тестовой выборке: для построенного с оправдывающей себя регуляризацией или вообще без неё? Почему?

    Если переобучение без регуляризации не превосходит дополнительный вес, который вносит регуляризация, то для алгоритма с оправдывающей себя регуляризацией.
    А если имеет место сильное переобучение, то для алгоритма вообще без регуляризации.

15. Что представляют собой метрики качества Accuracy, Precision и Recall?

    Precision and recall are then defined as:[6]

    $$\text{Precision} = \frac{tp}{tp + fp} \qquad \text{True negative rate} = \frac{tn}{tn + fp}$$

    $$\text{Recall} = \frac{tp}{tp + fn} \qquad \text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

    Здесь все введено для алгоритма классификации. Количество таких исходов, что:

    - tp - True Positive - алгоритм предсказал '+', и ответ тоже '+'
    - tn - True Negative - алгоритм предсказал '-', а ответ '-'
    - fp - False Positive - алгоритм предсказал '+', а ответ '-'
    - fn - False Negative - алгоритм предсказал '-', а ответ '+'

16. Что такое метрика качества AUC и ROC-кривая?

    ROC кривая - это график зависимости TPR от FPR sensitivity, recall, hit rate, or true positive rate (TPR)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

fall-out or false positive rate (FPR)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{N} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = 1 - \mathrm{TNR}$$

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

    Получаем множество:

    $$\{(FPR_i, TPR_i)\}_{i=1}^{l} \tag{9}$$

    полученное в результате следующего алгоритма:

    (a) Сортируем выборку x^l по значениям дискретной функции $f(x_i, \omega)$

    (b) Далее :

```
(FRP[0], TRP[0]) = (0, 0)
for i in range(1,l):
    if y_i == -1:
        (FPR[i], TPR[i]) = (FPR[i-1] + 1/l_negatives, TPR[i-1])
        (FPR[i], TPR[i]) = (FPR[i-1], TPR[i-1]  + 1/l_positives)
```

Здесь l_positives, l_negatives - число правильных и неправильных ответов в выборке соответственно.

2 Вероятностный смысл регуляризаторов

Покажите, что регуляризатор в задаче линейной классификации имеет вероятностный смысл априорного распределения параметров моделей. Какие распределения задают l1-регуляризатор и l2-регуляризатор?

- Пусть для параматерической модели задана плотность распределения

  $$p(x, y|\omega) \tag{10}$$

  $p(\omega)$ - априорная плотность.

  Совместное распределение выборки X^l и параметров рапределения ω по формуле условной вероятности - функция правдоподобия.

  $$p(X^l, \omega) = p(X^l|\omega) \cdot p(\omega) \tag{11}$$

  Логарифм функции правдоподобия:

  $$L(X^l, \omega) = \sum_{i=1}^{l} ln(p(x_i, y_i|\omega)) + ln(p(\omega)) \tag{12}$$

  Таким образом, слагаемое $ln(p(\omega))$ можно рассматривать как регуляризатор

- При $l1$ регуляризаторе априорное распределение параметров - распределение Лапласа с плотностью

  $$p(\omega) = \frac{1}{(2C)^n} e^{-\frac{1}{C}\sum\limits_{i=1}^{l}|\omega_i|} \tag{13}$$

  Потому что логарифм плотности:

  $$ln(p(\omega)) = -\frac{1}{C}\sum_{i=1}^{l}|\omega_i| + const \tag{14}$$

- При $l2$ регуляризаторе априорное распределение параметров - нормальное распределение с плотностью

$$p(\omega) = \frac{1}{(2\pi\sigma)^{n/2}} e^{-\frac{1}{2\sigma}\sum\limits_{i=1}^{l}\omega_i^2} \tag{15}$$

Логарифм плотности:

$$ln(p(\omega)) = -\frac{1}{2\sigma}\sum_{i=1}^{l}\omega_i^2 + const \tag{16}$$

3 SVM и максимизация разделяющей полосы

Покажите, как получается условная оптимизационная задача, решаемая в SVM из соображений максимизации разделяющей полосы между классами. Можно отталкиваться от линейно разделимого случая, но итоговое выражение должно быть для общего. Как эта задача сводится к безусловной задаче оптимизации?

4 Kernel trick

Придумайте ядро, которое позволит линейному классификатору с помощью Kernel Trick построить в исходном пространстве признаков разделяющую поверхность $x_1^2 + 2x_2^2 = 3$. Какой будет размерность спрямляющего пространства?

Рассмотрим ядро:

$$K(x,y) =< x, y >^2 = (x_1y_1 + x_2y_2)^2 = (x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 =< (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2y_2^2, \sqrt{2}y_1y_2) > \tag{17}$$

Получим отображение в спрямлённое пространство

$$H = R^3 : \psi(x_1, x_2) + (x_1^2, x_2^2, \sqrt{2}x_1x_2) \tag{18}$$

Линейная поверхность в H будет иметь вид

$$< (x_1^2, x_2^2, \sqrt{2}x_1x_2), (\omega_1, \omega 2, \omega 3) > +\omega_0 = \omega_1x_1^2 + \omega_2x_2^2 + \omega_3\sqrt{2}x_1x_2 + \omega_0 \tag{19}$$

Взяв $\omega = (1, 2, 0),\ \omega_0 = -3$, получим H в виде:

$$(19) = x_1^2 + x_2^2 - 3 = 0 \tag{20}$$

5 l_1-регуляризация

Покажите с помощью теоремы Куна-Таккера, что ограничение l_1-нормы вектора весов числом и добавление штрафа с его l_1-нормой приводят к построению одного и того же алгоритма. Можно считать, что регуляризатор добавляется по существу, т.е. меняет итоговый ответ по сравнению с оптимизационной задачей без регуляризатора.