Задание 4

Семен Федотов, 497 группа

Апрель, 2017

1 Знакомство с линейным классификатором

1 - 17

1. $a(x) = sign(< w, x > +w_0) = sign(\hat{y})$
2. Отступ это следующая величина $M(y_i, \hat{y_i}) = y_i \cdot \hat{y_i}$. Если отступ больше 0, то мы не ошиблись, если меньше - ошиблись. Если сильно положительный, то это говорит о сильной уверенности нашего алгоритма в этом ответе, а если он сильно отрицательный, то если наша модель хорошая, значит, этот элемент, скорее всего, выброс
3. Просто добавим еще одну компоненту для x, и у всех x она будет равна 1.
4. $Q(a, X) = \frac{1}{n}\sum[M_i \leq 0]$ У наилучшего он равен нулю, вед там нет ни одной ошибки , а следовательно все отступы меньше нуля
5. Возьмем веса равные нулю, тогда все индикаторы занулятся
6. $\frac{1}{n}\sum L(M)$, где M - отступ
7. Это функция, характеризующая величину ошибки на конкретном объекте. Хотим ее брать дифференцируемой, чтобы использовать градиентный спуск. При росте отступа, она невозрастает.
8. L = $[M \leq 0]$
9. Некоторая функция, зависящая от весов алгоритма.(В статистическом смысле ее можно понимать как априорное распределение на весах). $L_1 \rightarrow laplace, \quad L_2 \rightarrow multinomial \quad normal$ Первая - сумма модулей остатков, а вторая - сумма квадратов остатков.
10. Регуляризация часто помогает бороться с переобучением, например, в случае, если у нашей модели получились очень большие веса(она сильно подстроилась под тренировочную выборку), а когда к нам придет новый элемент мы получим ужасный ответ, это и есть обобщающая способность.
11. Если мы выйдем из него, то значение функции риска сильно скакнет
12. При выходе за границы, функционал риска будет расти
13. С регуляризатором, ведь это неотрицательная функция и она увеличивает ее значение
14. Все зависит от ситуации, может произойти и то, и то. Мы можем сильно переобучиться без регуляризатора и получить гигантское значение функции риска, но также и с регуляризатором, если все же он не сильно уж и нужен в нашей ситуации
15. Accuracy = $\frac{\sum\limits_{i=1}^{n}[M_i > 0]}{n}$, Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$
16. AUC - area under curve. ROC(Receiver operating characteristic) - кривая построенная в осях TPR, FPR, где TPR = $\frac{TP}{TP+FN}$ = Precision, FPR = $\frac{FP}{FP+TN}$
17. Ну, нужно попереберать границы(threshold), по которым мы будем определять к какому все же классу отнести очередной элемент. будем перебирать просто границы как какое-то число между двумя соседними элементами в выборке.

2 Вероятностный смысл классификаторов

Как я говорил ранее, L_1 и L_2 регуляризации задают априорные распределения Лапласа и Многомерное Нормальное, соответственно. Как это показать? Ну мы собираемся минимизировать функционал ошибки, который в нашем случае равен $L = \sum L_i + R(w)$. Хотим считать, что это все логарифм правдоподобия какой-то выборки, т.е $L = \sum lnp(y_i, x_i \mid \omega) + lnp(\omega, \lambda) \overset{\max}{\rightarrow}$. Отсюда видно что, $R(\omega)$ задает априорное распределение: $e^R = p(\omega, \lambda)$

3 SVM + Maximize stripe width

У нас выборка линейно разделима, а что это значит? $Q(w) = \sum[M \leq 0] = 0$. Круто, но нам нужно расположить полосу так, чтобы расширить разделяющуюся полосу. Изменим веса(домножим на какое-то число), чтобы минимальное значение отступа было равно 1. Так как мы хотим расширить полосу, то на ее границах должны лежать точки разных классов, обозначим их (x_-, x_+), ну тогда ширину полосы мы можем найти следующим образом(возьмем просто проекцию вектора, соединябщего эти точки на вектор весов): $< x_+ - x_-, \frac{w}{||w||} >= \frac{<x_+,w>-<x_-,w>}{||w||}$ / Так как мы провели ту нормировку, то оступы везде равны $1/ = \frac{2}{||w||}$. Раз мы хотим ее максимищровать, то это то же самое, что минимизировать $\frac{1}{2}||w||^2$ и еще нужно не забыть условие, которого мы добились с помощью нормировки(минимум по всем значения отступа больше 1). В случае линейно неразделимой выборки у нас отсуп может и не быть больше 1, а поэтому нужно добавить новые переменные $\xi_i \geq 0$. Тогда получим следующую оптимизационную задачу:

$$\frac{1}{2}||w||^2 + \lambda \cdot \sum \xi_i$$

$$M_i \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

4 Kernel Trick

Это эллипс, возьмем квадратичное ядро: $K(x,y) =< x, y >^2= (x_1y_1 + x_2y_2)^2 = (x_1y_1)^2 + 2x_1x_2y_1y_2 + (x_2y_2)^2 =< (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y1_y1) >$. Как видно, это пространство размерности 3