Знакомство с линейным классификатором

1. $f(x) =< w, x > +w_0$

   $class(x) = 1$, если $f(x) > 0$

   $class(x) = -1$, иначе

2. Отступом на объекте называется $M(x) = class_{actual}(x) \cdot f(x)$

   $M(x) > 0$ если $class(x) = class_{actual}(x)$

   $M(x) < 0$ если $class(x) \neq class_{actual}(x)$

3. Всем объектам добавляют дополнительный признак равный 1.

4. $$Q(w) = \sum_{i=1}^{l} [M_i(w) < 0]$$
   $$w_{best} = argmin(Q(w))$$

5. $w = (0, ..., 0)$

6. $$\widetilde{Q}(w) = \sum_{i=1}^{l} L(M_i(w))$$

   где L - функция потерь

7. Функция потерь нужна чтобы аппроксимировать функционал эмпирического риска и при этом иметь возможность применять применять методы оптимизации для ее минимизации. Такие функции обычно возрастают при $M \to -\inf$ и стремятся к нулю при $M \to +\inf$, и, судя по всему, в нуле обычно равны 1.

8. $V(M) = (1 - M)_+$

9. Регуляризация - это способ уменьшить переобучение линейной модели, штрафующий за большие по модулю веса в модели. Основные регуляризаторы это:

   - сумма модулей весов (L_1 регуляризатор)
   - сумма квадратов весов (L_2 регуляризатор)

10. Переобучение само по себе означает подстроение модели к обучающей выборке, то есть уменьшение обобщающей способности. А поскольку резуляризация препятствует переобучению, то она улучшает обобщающую способность.

Повторение: метрики качества

1. Accuracy: доля правильных ответов классификации

   Precision: TruePositive / (TruePositive + FalsePositive)

   Recall: TruePositive / (TruePositive + FalseNegative)

2. $TPR = \frac{TruePositive}{TruePositive+FalseNegative}$

   $FPR = \frac{FalsePositive}{FalsePositive+TrueNegative}$

   ROC кривая - график зависимости TPR от FPR. Мера AUC - площадь под ROC.

3. Постоить ROC кривую можно следующим образом:

   (a) По известным ответам посчитать число объектов обоих классов m_-, m_+

   (b) Упорядочить выборку по убыванию $probability(x)$

   (c) Установить начальное значение (0, 0)

   (d) Перебирем все объекты выборки

   - если очередной объект принадлежит классу -1 то сместимся вправо: $FPR := FPR + \frac{1}{m_-}$
   - иначе вверх: $TPR := TPR + \frac{1}{m_+}$