

Машинное обучение: задание 4.

3.1. Знакомство с линейными классификатором.

1. Бинарный линейный классификатор

$$a(x) = \text{sign}(f(x)), \quad f(x) = w_0 + (w, x)$$

2. Отступ алгоритма на объекте.

$$M_i = y_i f(x_i) = \begin{cases} > 0, & \text{класс угадан верно} \\ < 0, & \text{--- неверно} \end{cases}$$

3. Чтобы из $a(x) = \text{sign}(+w_0 + (w, x))$ получить классификатор вида $a(x) = \text{sign}((w, x))$, нужно ввести элемент x_0 , и добавить к вектору x , а w_0 добавить к w .

4. Запись функционала эмпирического риска через отступы:

$$Q(x) = \frac{1}{m} \sum_{i=1}^m I(M_i \leq 0)$$

Для "наилучшего" алгоритма классификации: $Q(x) = 0$

5. Если $Q(x) = \frac{1}{m} \sum_{i=1}^m I(M_i < 0)$, то при

$$w_i = 0 \quad \forall i \quad \rightarrow \quad I(M_i < 0) = 0 \Rightarrow Q(x) = 0.$$

6. Функционал аппроксимированного эмпирического риска, если выбрать функцию потерь $L(M)$.

$$\tilde{Q}(w) = \sum_{i=1}^m L(M_i(w))$$

7. Функция потерь характеризует ошибку алгоритма $L(f(x_i), y_i)$. ϕ -я потерь неограниченная, невозрастающая.

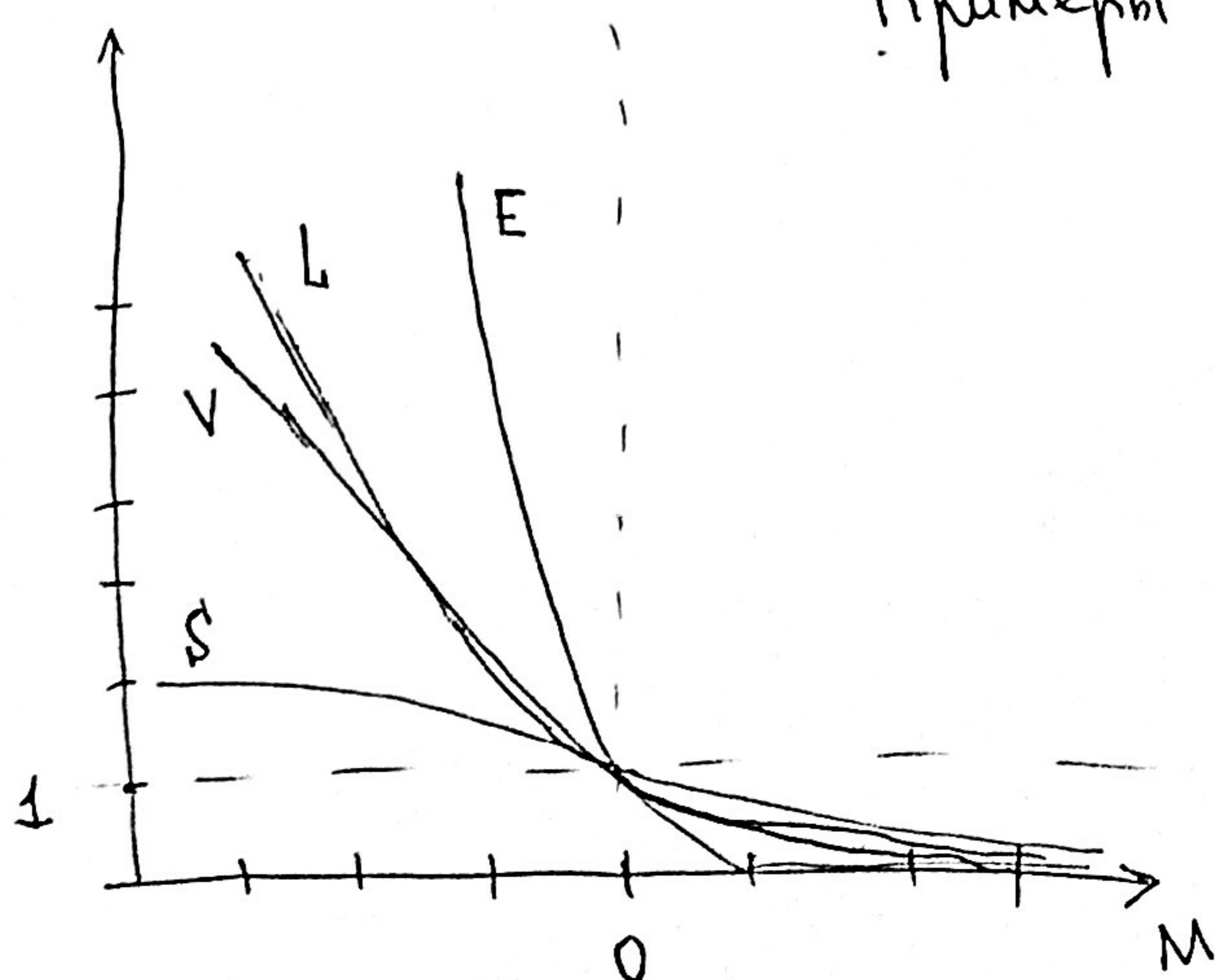
Примеры:

$$E(M) = e^{-M}$$

$$L(M) = \log_2(1 + e^{-M})$$

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

$$S(M) = 2(1 + e^{-M})^{-1}$$



8. Пример неплавкой ф-и потерь

$$L(M) = (1 - M)_+$$

9. Регуляризация ; регуляризаторы

Регуляризация - добавление ограничений на коэффициенты, т.к. большие коэффициенты делают возможным сильное изменение величины при небольшом изменении признаков.

l1 - регуляризация

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{k=1}^m |w_k| \rightarrow \min$$

l2 - регуляризация

$$\sum_{i=1}^l L(M_i) + \gamma \sum_{k=1}^m w_k^2 \rightarrow \min$$

10. Как связаны переобучение и обобщающая способность алгоритма?

Обобщающая способность алгоритма: говорит, что алгоритм обучен не обладает способностью к обобщению, если вероятность ошибки на тестовой выборке достаточно мала или предсказуема, т.е. не сильно отличается от ошибки на обучающей выборке.

При переобучении вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Возникает при использовании избыточно сложных моделей.

Регуляризация ограничивает веса, эти примерно одного порядка. Переобучение может быть, если нет регуляризации, и эти какие-то веса доминируют.

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

$$\hat{Q}(w) = \sum_{i=1}^l L(M_i(w)) \rightarrow \min$$

При острых минимумах $\hat{Q}(w)$ будет сильно увеличиваться вес w_k , будет переобучение.

12. Что делает регуляризацию с аппроксимированным риском функцией параметров алгоритма?

Функция будет увеличиваться, когда веса будут близки к определенным допустимым значениям.

13. Для какого алгоритма классификации функционал \hat{Q} - $\hat{\rho}$ будет принимать бо́льшее значение на обучающей выборке: построенного с регуляризацией или без неё?

Для алгоритма с регуляризацией

$$\hat{Q} = \sum_{i=1}^l L(M_i) \rightarrow \min \quad \leftarrow \text{без регуляризации}$$

При регуляризации добавляется ещё одно неотрицательное слагаемое. К тому же, без регуляризации алгоритм бы больше переобучился на обучающей выборке, стараясь подобрать веса так, чтобы минимизировать \hat{Q} , с регуляризатором мы даём ограничение на веса и на обучающей выборке алгоритм уже не сможет так сильно подстроиться под неё.

14. Для какого алгоритма классификации функционал риска будет принимать бо́льшее значение на тестовой выборке: для построенного с оптимальной себе регуляризацией или бо́льше без неё?

На тестовой выборке может быть иначе.

Без регуляризации алгоритм может сильно переобучиться на обучающей выборке и на тестовой ~~не~~ дать результат хуже, чем мог бы без регуляризатора, ограничивающего переобучение.

С другой стороны тестовая выборка может быть настолько похожа на обучающую, что переобучение, наоборот, даст меньшую функциональную риску, нежели тем при использовании регуляризации.

15. Accuracy - доля правильных ответов при классификации.

Precision - точность

Recall - полнота

$$\text{Precision} = \frac{TP}{TP + FP} ; \quad \text{Recall} = \frac{TP}{TP + FN}$$

16. ROC-кривая - график, позволяющий оценить качество бинарной классификации, отображает соотношение между TPR и FPR.

$$TPR = \frac{TP}{TP + FN} ; \quad FPR = \frac{FP}{FP + TN}$$

ROC-AUC - площадь под ROC-кривой. Чем выше AUC, тем качественнее классификатор. Значение 0,5 демонстрирует непригодность бинарного метода классификации.

47. Как настроить ROC - кривую

1. Выделить кон-бо представителей классов +1 и -1 в выборке:

$$m_- = \sum_{i=1}^m I(y_i = -1)$$

$$m_+ = \sum_{i=1}^m I(y_i = +1)$$

2. Упорядочить выборку X^m по убыванию значений $f(x_i, w)$ ($w = \text{фикс.}$)

3. Установить начальное значение ROC - кривой:

$$FPR_0 = 0$$

$$TPR_0 = 0$$

4. Для всех $i = \overline{1..m}$

если $y_i = -1$, то сместить вправо 1 шаг вправо

$$FPR_i = FPR_{i-1} + \frac{1}{m_-}$$

$$TPR_i = TPR_{i-1}$$

если $y_i = 1$, то вправо 1 шаг влево

$$FPR_i = FPR_{i-1}$$

$$TPR_i = TPR_{i-1} + \frac{1}{m_+}$$

Понятие $FPR_i, TPR_i \quad i = \overline{0, m}$ — надеждочательность точек ROC - кривой.

□.

Задача 3.6 Повторение : методы настройки
соответствия с 3.1 15-17.

3.2. Вероятностный смысл регуляризаторов

$$Q = \sum_{i=1}^l L(y_i, f(x_i)) + \gamma V(w) \rightarrow \min_w$$

↑ регуляризатор

$$\sum_{i=1}^l -L(y_i, f(x_i)) - \gamma V(w) \rightarrow \max_w$$

$$\sum_{i=1}^l \ln \exp \{-L(y_i, f(x_i))\} + \ln \{\exp [-\gamma V(w)]\} \rightarrow \max_w$$

$$\underbrace{\exp [-\gamma V(w)]}_{P(w)} \cdot \prod_{i=1}^l \underbrace{\exp [-L(y_i, f(x_i))]}_{P(x_i, y_i | w_i)} \rightarrow \max_w$$

Регуляризатор в задаче нн. классификации имеет вероятностный смысл эмпирического распределения параметров подъем.

1) Пусть $w \in \mathbb{R}^n$ имеет n -мерное гауссовское распределение:

$$p(w, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\|w\|^2}{2\sigma}\right), \quad \|w\|^2 = \sum_{j=1}^n w_j^2.$$

$$\text{Логарифмруя: } -\ln p(w; \sigma) = \frac{1}{2\sigma} \|w\|^2 + \text{const}$$

$\Rightarrow l_2$ - гауссовский регуляризатор

2) Пусть $w \in \mathbb{R}^n$ имеет n -мерное распределение Лапласа:

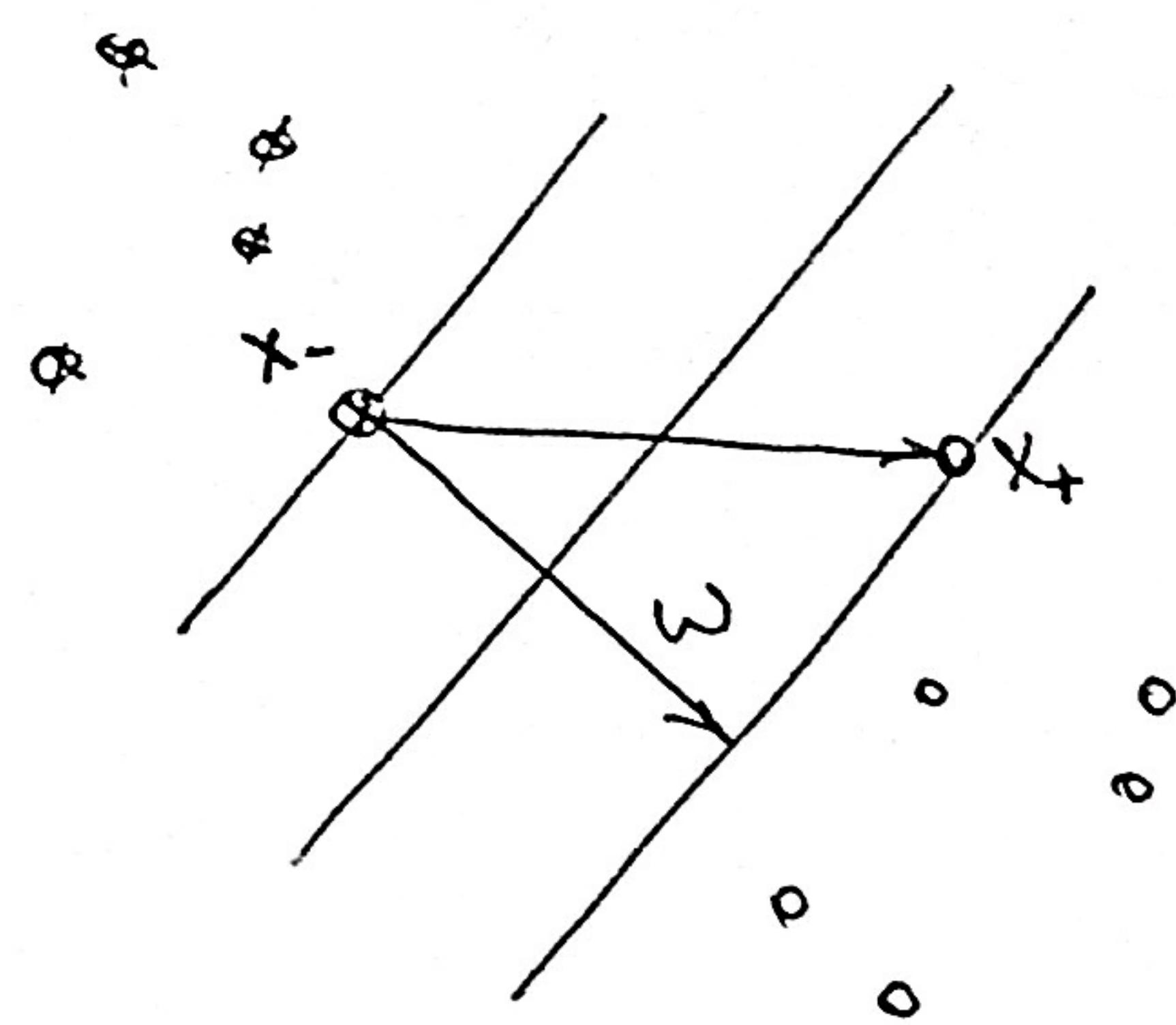
$$p(w, c) = \frac{1}{(2c)^n} \exp\left(-\frac{\|w\|_1}{c}\right), \quad \|w\|_1 = \sum_{j=1}^n |w_j|$$

$$\text{Логарифмруя: } -\ln p(w, c) = \frac{1}{c} \sum_{j=1}^n |w_j| + \text{const}(w)$$

$\Rightarrow l_1$ - лапласовский регуляризатор.

□.

3.3. SVM и максимизация разделяющих плоскостей.



Разделяющие плоскости

$$a(x) = \text{sign}(\langle w, x \rangle - w_0)$$

$$\min_{i=1 \dots l} y_i (\langle w, x_i \rangle - w_0) = 1$$

Минимизация разделяющих плоскостей

$$\left\langle (x_+ - x_-), \frac{w}{\|w\|} \right\rangle =$$

$$= \frac{\langle w, x_+ \rangle - \langle w, x_- \rangle}{\|w\|} = \frac{(w_0 + 1) - (w_0 - 1)}{\|w\|} = \frac{2}{\|w\|}$$

Максимизация разделяющих плоскостей

$$\begin{cases} \langle w, w \rangle \rightarrow \min \\ y_i (\langle w, x_i \rangle - w_0) \geq 1 \quad \forall i \end{cases}$$

Случай линейно неразделимых выборок

$$\begin{cases} \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^l \xi_i \rightarrow \min \\ y_i (\langle w, x_i \rangle - w_0) \geq 1 - \xi_i \\ \xi_i \geq 0 \quad \forall i \end{cases}$$

Оптимизационная задача в SVM

Безусловная оптимизационная задача в SVM

$$\frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^l \xi_i \quad M_i = y_i (\langle w, x_i \rangle - w_0)$$

$$\xi_i \geq 0$$

$$\xi_i \geq 1 - M_i \Rightarrow \xi_i = \max \{0, 1 - M_i\} = (1 - M_i)_+$$

$$\sum_{i=1}^l \xi_i \rightarrow \min$$

$$Q(w, w_0) = \sum_{i=1}^l (1 - M_i(w, w_0))_+ + \frac{1}{2C} \|w\|^2 \rightarrow \min_{w, w_0}$$

□.

3.4. Kernel trick -

▲ способ создания нелинейного классификатора на основе которого лежит переход от линейных произведений к произвольным выражениям.

Понятие ядра
Рассмотрим $k(x, x') = (x, x')$
 $= (x_1 x'_1)^2 + 2(x_1 x'_1)(x_2 x'_2) + (x_2 x'_2)^2 =$
 / можно представить как / $= (x_1^2, \sqrt{2}x_1 x_2, x_2^2)$ $\begin{pmatrix} x_1^{12} \\ \sqrt{2}x_1 x_2 \\ x_2^{12} \end{pmatrix}$

Получаем отображение из (x, x') в $(x^2, x^{12}, x x')$

Мы видим в исходном виде признаков построить разделяющую поверхность: $x_1^2 + 2x_2^2 = 3$.

В сформированном пространстве:

$$\langle x^*, w \rangle - w_0 = 0$$

$$\downarrow$$

$$(x_1^2, x_2^2, \sqrt{2}x_1 x_2) \cdot \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} - w_0 = 0$$

Получим лин: $w_0 = 3, w_1 = 1, w_2 = 2, w_3 = 0$.

$$x_1^2 + 2x_2^2 = 3$$

Размерность сформированного нп-ла = 3.

□.

3.5 l_1 - регуляризация

$$\left\{ \begin{array}{l} \sum_{i=1}^m L(x_i, w, y_i) \rightarrow \min_w \\ \sum_{j=1}^m |w_j| \leq \bar{\tau} \end{array} \right. \quad \leftarrow \text{ограничение } l_1\text{-нормы вектора весов модели (1)} \quad (1)$$

Теорема Каруша - Куна - Таккета

$$\left\{ \begin{array}{l} f(x) \rightarrow \min \\ g(x) \leq 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \nabla_x L(x, \mu) = \nabla_x (f(x) + \mu g(x)) = 0 \\ \mu g(x) = 0 \\ \mu \geq 0 \end{array} \right.$$

$$\sum_{j=1}^m |w_j| \leq \bar{\tau} \iff \sum_{j=1}^m |w_j| - \bar{\tau} \leq 0$$

но т. к. кт:

$$(1) \Rightarrow \left\{ \begin{array}{l} \sum_{i=1}^m L(x_i, w, y_i) + \mu \left(\sum_{j=1}^m |w_j| - \bar{\tau} \right) \rightarrow \min \\ \mu \left(\sum_{j=1}^m |w_j| - \bar{\tau} \right) = 0 \\ \mu \geq 0 \end{array} \right. \quad \text{const.}$$

т.е. получаем

$$\left\{ \begin{array}{l} \sum_{i=1}^m L(x_i, w, y_i) + \mu \sum_{j=1}^m |w_j| \rightarrow \min \\ \mu \geq 0 \\ \mu = 0 \text{ или } \sum_{j=1}^m |w_j| = \bar{\tau} \end{array} \right.$$

добавление штрафа (2)
с и/o l_1 -нормой

т.е. (1) и (2) приводят к получению одицко и то же
алгоритма.

□.