Машинное обучение, ФКН ВШЭ
Линейные модели классификации

Задача 1. Пусть даны выборка выборка X, состоящая из 8 объектов, и классификатор $b(x)$, предсказывающий оценку принадлежности объекта положительному классу. Предсказания $b(x)$ и реальные метки объектов приведены ниже:

$$
\begin{aligned}
b(x_1) &= 0.1, \quad y_1 = +1,\\
b(x_2) &= 0.8, \quad y_2 = +1,\\
b(x_3) &= 0.2, \quad y_3 = -1,\\
b(x_4) &= 0.25, \quad y_4 = -1,\\
b(x_5) &= 0.9, \quad y_5 = +1,\\
b(x_6) &= 0.3, \quad y_6 = +1,\\
b(x_7) &= 0.6, \quad y_7 = -1,\\
b(x_8) &= 0.95, \quad y_8 = +1.
\end{aligned}
$$

Построим ROC-кривую вручную по следующим пунктам:

1. Отсортировать данные по невозрастанию оценки вероятности принадлежности к первому классу:

$$
\begin{aligned}
b(x_8) &= 0.95, \quad y_8 = +1\\
b(x_5) &= 0.9, \quad y_5 = +1\\
b(x_2) &= 0.8, \quad y_2 = +1\\
b(x_7) &= 0.6, \quad y_7 = -1\\
b(x_6) &= 0.3, \quad y_6 = +1\\
b(x_4) &= 0.25, \quad y_4 = -1\\
b(x_3) &= 0.2, \quad y_3 = -1\\
b(x_1) &= 0.1, \quad y_1 = +1
\end{aligned}
$$

2. Разметить горизонтальную ось (FPR) на количество объектов истинного отрицательного класса. Разметить вертикальную ось (TPR = Recall) на количество объектов истинного положительного класса.

3. Начинать строить кривую с точки $(0, 0)$. Идти по таблице сверху внизу, если мы встречаем среди истиных меток 1, то сдвигаемся на одну "клетку"вверх, если мы встречаем 0, то вправо. Если мы встречаем пару различных истиных меток 0/1, но при этом одинаковое значение предсказанной вероятности: строим кривую по диагонали.

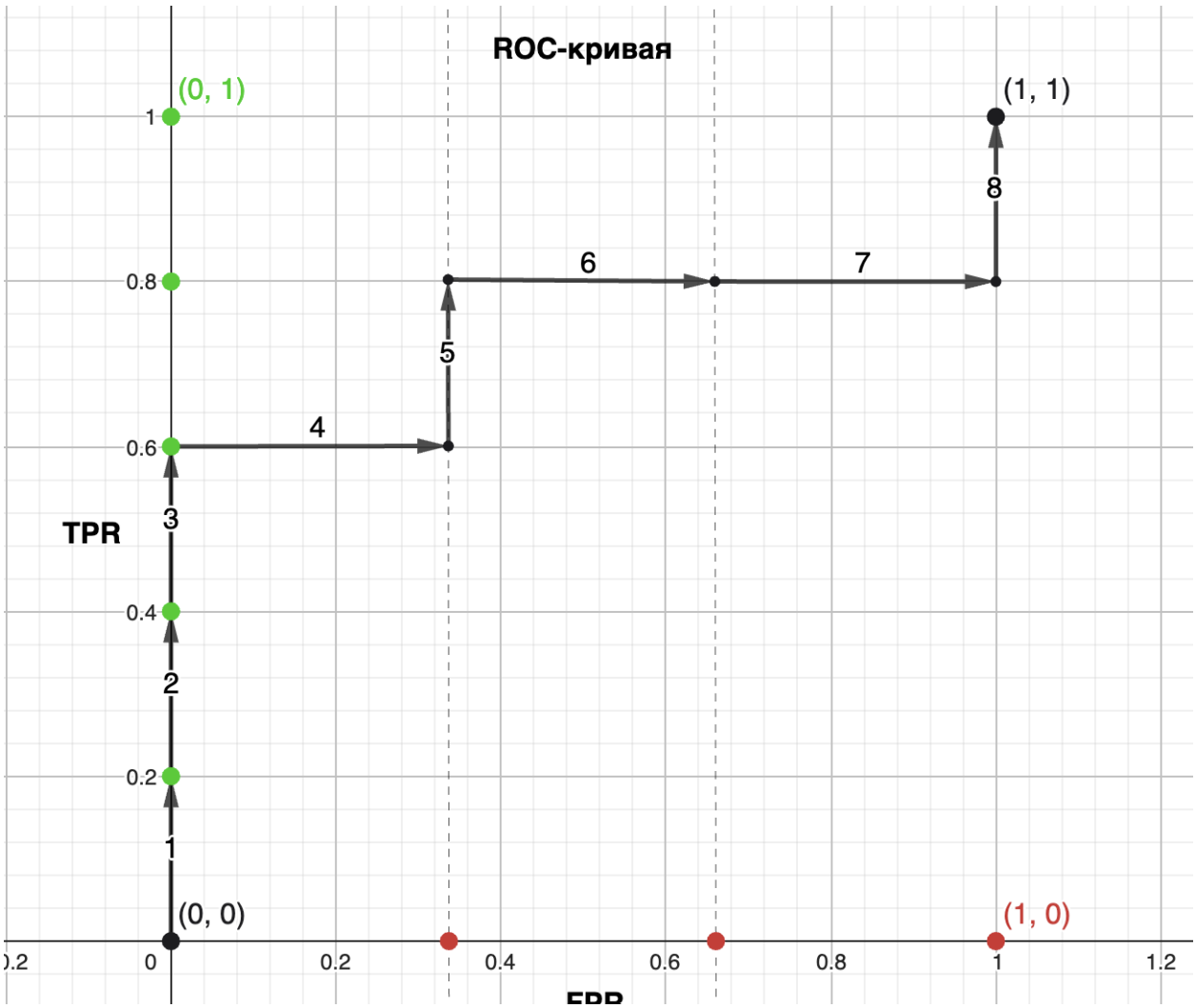

Рис. 1: ROC-curve

4. Кривая заканчивается в точке $(1, 1)$

Затем необходимо посчитать площадь под кривой:

$$AUC = 0.6 \cdot \frac{1}{3} + 0.8 \cdot \frac{2}{3} \approx 0.73$$

Проверка:

```
from sklearn.metrics import roc_auc_score
```

```
y_true = [1, 1, 1, 0, 1, 0, 0, 1]
y_pred = [0.95, 0.9, 0.8, 0.6, 0.3, 0.25, 0.2, 0.1]

roc_auc_score(y_true, y_pred)
```

```
0.7333333333333334
```

Рис. 2: sklearn ROC-AUC

Задача 2. Пусть дана некоторая выборка X и классификатор $b(x)$, возвращающий в качестве оценки принадлежности объекта x положительному классу 0 или 1 (а не некоторое вещественное число, как предполагалось на семинарах).

1. Постройте ROC-кривую для классификатора $b(x)$ на выборке X.
   Это будет ломаная, проходящая через точки $(0,0), (FPR, TPR), (1,1)$

2. Покажите, что AUC-ROC классификатора $b(x)$ на выборке X может быть выражен через долю правильных ответов и полноту классификатора $a(x;t)$, получающегося при выборе некоторого порога $t \in (0;1)$. Помимо указанных величин в формулу могут входить только величины ℓ_-, ℓ_+, ℓ (количество отрицательных, положительных и общее количество объектов в выборке X соответственно).

   Площадь под ломаной, проходящей через $(0,0)$, (TPR, FPR), $(1,1)$ будет равна площади двух треугольников и одного прямоугольника:

$$\begin{aligned} ROC - AUC = &\frac{1}{2}(FPR \cdot TPR) + \\ &+ \frac{1}{2}((1 - FPR)(1 - TPR)) + \\ &+ \frac{1}{2}(2 \cdot TPR \cdot (1 - FPR)) = \\ &= \frac{1}{2}(TPR + (1 - FPR)) \end{aligned}$$

   Выразим через требуемые величины:

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{l_+} = Recall$$

$$TP = Recall \cdot l_+$$

$$Accuracy = \frac{TP + TN}{l} = \frac{l_- - FP + TP}{l}$$

$$FP = TP + l_- - Accuracy \cdot l = Recall \cdot l_+ + l_- - Accuracy \cdot l$$

$$FPR = \frac{FP}{l_-}$$

$$l_+ = l - l_-$$

   Подставим:

$$ROC - AUC = \frac{1}{2}\left(TPR + (1 - FPR)\right) = Recall + \frac{l \cdot (Accuracy - Recall)}{2l_-}$$

3. Покажите, что в случае сбалансированной выборки ($\ell_- = \ell_+$) AUC-ROC классификатора $b(x)$ на выборке X совпадает с долей правильных ответов классификатора при выборе некоторого порога $t \in (0;1)$.

   Т.к. $\ell_- = \ell_+$, то $l = 2l_-$, тогда из предыдущего пункта:

$$ROC - AUC = Recall + \frac{2l_- \cdot (Accuracy - Recall)}{2l_-} = Accuracy$$

Задача 3. В анализе данных для сравнения среднего значения некоторой величины у объектов двух выборок часто используется критерий Манна–Уитни–Уилкоксона, основанный на вычислении U-статистики.

Пусть у нас имеется выборка X и классификатор $b(x)$, возвращающий оценку принадлежности объекта x положительному классу. Тогда вычисление U-статистики для подвыборки X, состоящей из объектов положительного класса, производится следующим образом: объекты обеих выборок сортируются по неубыванию значения $b(x)$, после чего каждому объекту в полученном упорядоченном ряду $x_{(1)}, \ldots, x_{(\ell)}$ присваивается ранг — номер позиции $r_{(i)}$ в ряду (начиная с 1, при этом для объектов с одинаковыми значением $b(x)$ в качестве ранга присваивается среднее значение ранга для таких объектов). Тогда U-статистика для объектов положительного класса равна:

$$U_+ = \sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} r_{(i)} - \frac{\ell_+(\ell_+ + 1)}{2}.$$

Покажите, что для значения AUC-ROC классификатора $b(x)$ на выборке X и U-статистики верно следующее соотношение:

$$\text{AUC} = \frac{U_+}{\ell_-\ell_+}.$$

Одна из формул для вычисления AUC: вероятность того, что выход классификатора для объекта положительного класса будет больше, чем выход классификатора для объекта отрицательного класса

$$\text{AUC} = P(b(x^+) > b(x^-)),$$

Если есть объекты с одинаковыми значениями $b(x)$, то:

$$\text{AUC} = P(b(x^+) > b(x^-)) + \frac{1}{2}P(b(x^+) = b(x^-)).$$

Теперь рассмотрим U-статистику для положительного класса:

$$U_+ = \sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} r_{(i)} - \frac{\ell_+(\ell_+ + 1)}{2},$$

где $r_{(i)}$ — ранг i-го объекта в общем упорядоченном ряду, ℓ_+ — количество объектов положительного класса. Сумма

$$\sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} r_{(i)}$$

— это сумма рангов объектов положительного класса.

$$\frac{\ell_+(\ell_+ + 1)}{2}$$

— это сумма рангов, которую имели бы объекты положительного класса, если бы они все были в начале отсортированного списка, то есть , сумма первых l_+ членов натурального ряда.

Рассмотрим, сколько пар объектов (x^+, x^-) таких, что $b(x^+) > b(x^-)$. Каждый объект

положительного класса с рангом $r_{(i)}$ имеет перед собой $(r_{(i)}-1)$ объектов. Из них $(r_{(i)}-i)$ объектов относятся к отрицательному классу (поскольку i-й по счёту объект положительного класса может иметь перед собой максимум $(i-1)$ объектов положительного класса).
Таким образом, общее число пар (x^+, x^-) таких, что $b(x^+) > b(x^-)$, равно:

$$\sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} (r_{(i)} - i).$$

Преобразуем это выражение:

$$\sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} (r_{(i)} - i) = \sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} r_{(i)} - \sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} i.$$

Вторая сумма — это сумма первых ℓ_+ натуральных чисел:

$$\sum_{i=1}^{\ell_+} i = \frac{\ell_+(\ell_+ + 1)}{2}.$$

Таким образом,

$$\sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} (r_{(i)} - i) = \sum_{\substack{i=1 \\ y_{(i)}=+1}}^{\ell} r_{(i)} - \frac{\ell_+(\ell_+ + 1)}{2} = U_+.$$

Теперь вычислим AUC:

$$\text{AUC} = \frac{\text{число пар } (x^+, x^-) \text{ таких, что } b(x^+) > b(x^-) + \frac{1}{2} \cdot \text{число пар с } b(x^+) = b(x^-)}{\text{общее число пар } (x^+, x^-)}.$$

Общее число пар (x^+, x^-) равно $\ell_+ \cdot \ell_-$.
Если нет объектов разного класса с одинаковым выходом классификатора, то число пар с $b(x^+) > b(x^-)$ равно U_+.
Для объектов с одинаковыми значениями $b(x)$ присваивается среднее значение ранга. Это эквивалентно учёту половины пар с равными значениями.
Таким образом,

$$\text{AUC} = \frac{U_+}{\ell_+ \cdot \ell_-},$$

Задача 4. Позволяет ли предсказывать корректные вероятности экспоненциальная функция потерь $L(y, z) = \exp(-yz)$?

Позволяет, как минимум, алгоритм AdaBoost использует экспоненциальную функцию потерь, выходы которой можно трактовать как вероятности принадлежности к классу.
Покажем это:
Пусть $y \in \{0, 1\}$ - это истинные метки классов, а z - выход модели. Функция потерь позволяет предсказывать корректные вероятности тогда, когда она минимизируется только выходом модели, соответствующим отношению вероятности классов.
Рассчитаем мат.ожидание:

$$\mathbb{E}[L(y, z) \mid x] = P(y = +1 \mid x) \cdot e^{-z} + P(y = -1 \mid x) \cdot e^{z}.$$

Пусть: $\eta = P(y = +1 \mid x)$, тогда $P(y = -1 \mid x) = 1 - \eta$ Подставим, продифференцируем и найдем минимум по z:

$$\frac{d}{dz}\mathbb{E}[L(y,z)] = -\eta \cdot e^{-z} + (1-\eta) \cdot e^{z} = 0$$

$$z = \frac{1}{2}\ln\frac{\eta}{1-\eta} = \frac{1}{2}\ln\frac{P(y=+1 \mid x)}{P(y=-1 \mid x)}$$

Таким образом, если преобразовать выход модели $z \to 2z$, то экспоненциальная функция потерь действительно будет предсказывать корректные вероятности, их можно получить с помощью обратного преобразования:

$$P(y = +1 \mid x) = \frac{1}{1 + e^{-2z}}$$

Задача 5. Рассмотрим постановку оптимизационной задачи метода опорных векторов для линейно разделимой выборки:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min_{w,b}, \\ y_i(\langle w, x\rangle + b) \geq 1, \quad i = \overline{1,\ell}, \end{cases}$$

а также её видоизменёный вариант для некоторого значения $t > 0$:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min_{w,b}, \\ y_i(\langle w, x\rangle + b) \geq t, \quad i = \overline{1,\ell}. \end{cases}$$

Покажите, что разделяющие гиперплоскости, получающиеся в результате решения каждой из этих задач, совпадают.

Введем переменные $w^{'} = w/t$ и $b^{'} = b/t$ Подставляя эти переменные во вторую задачу, получаем:

$$\begin{cases} \frac{1}{2}\|t \cdot w'\|^2 \to \min_{w',b'}, \\ y_i(\langle t \cdot w', x_i\rangle + t \cdot b') \geq t, \quad i = \overline{1,\ell} \end{cases}$$

Упрощая:

$$\begin{cases} \frac{1}{2} \cdot t^2 \cdot \|w'\|^2 \to \min_{w',b'}, \\ t \cdot y_i(\langle w', x_i\rangle + b') \geq t, \quad i = \overline{1,\ell} \end{cases}$$

Сокращая $t > 0$ в ограничениях и константный множитель $\frac{t^2}{2}$, т.к. он никак не влияет на положение минимума, получаем:

$$\begin{cases} \frac{1}{2} \cdot \|w'\|^2 \to \min_{w',b'}, \\ y_i(\langle w', x_i\rangle + b') \geq 1, \quad i = \overline{1,\ell} \end{cases}$$

Эта задача идентична первоначальной задаче, но с переменными $w^{'}$ и $b^{'}$.
Таким образом, если w^* и b^* являются оптимальными решениями первой задачи, то оптимальными параметрами второй задачи будут $w^{'*} = t \cdot w^*$ и $b^{'*} = t \cdot b^*$. // Разделяющая гиперплоскость определяется уравнением

$$\langle w, x\rangle + b = 0$$

Для второй задачи:

$$\langle t \cdot w, x \rangle + (t \cdot b) = 0$$

Т.к. $t > 0$, это эквивалентно

$$\langle w, x \rangle + b = 0$$

Это уравнение задает ту же гиперплоскость, что и в первой задаче. Следовательно, разделяющие гиперплоскости, получающиеся при решении первой и второй задач, совпадают.

Задача 6. Пусть мы решили двойственную задачу SVM и получили оптимальные значения $(\lambda_1, \ldots, \lambda_\ell)$, где $\lambda_5 = C/3$, $\lambda_2 = 0$. Выразите оптимальное значение порога b для прямой задачи через найденное решение $(\lambda_1, \ldots, \lambda_\ell)$ двойственной задачи.

В прямой задаче SVM мы оптимизируем:

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \to \min\limits_{w,b,\xi}, \\ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad i = 1, \ldots, \ell, \\ \xi_i \geq 0, \quad i = 1, \ldots, \ell \end{cases}$$

В двойственной задаче мы находим множители Лагранжа λ_i. По условиям Каруша – Куна – Таккера оптимальные значения параметров прямой и двойственной задач связаны следующими соотношениями:

1. $w = \sum\limits_{i=1}^{\ell} \lambda_i y_i x_i$;
2. $\sum\limits_{i=1}^{\ell} \lambda_i y_i = 0$;
3. Для каждого i выполняется одно из:
   - $\lambda_i = 0$ и $y_i(\langle w, x_i \rangle + b) \geq 1$;
   - $0 < \lambda_i < C$ и $y_i(\langle w, x_i \rangle + b) = 1$;
   - $\lambda_i = C$ и $y_i(\langle w, x_i \rangle + b) \leq 1$.

Для нахождения оптимального порога b используем второе условие: если $0 < \lambda_i < C$, то соответствующий объект лежит точно на границе зазора, и для него $y_i(\langle w, x_i \rangle + b) = 1$. По условию задачи $\lambda_5 = C/3$, что означает $0 < \lambda_5 < C$. Следовательно, объект x_5 является опорным вектором, лежащим точно на границе зазора, и мы можем использовать его для определения b:

$$y_5(\langle w, x_5 \rangle + b) = 1$$

Выразим b из этого уравнения:

$$b = \frac{1}{y_5} - \langle w, x_5 \rangle$$

Подставляя выражение для w:

$$b = \frac{1}{y_5} - \left\langle \sum_{i=1}^{\ell} \lambda_i y_i x_i, x_5 \right\rangle = \frac{1}{y_5} - \sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x_5 \rangle$$

Таким образом:

$$b = \frac{1}{y_5} - \sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x_5 \rangle$$

где y_5 — метка класса для объекта x_5, а $\langle x_i, x_5 \rangle$ — скалярное произведение векторов x_i и x_5.

Задача 7. Вычислите градиент $\frac{\partial}{\partial w} L(x, y; w)$ логистической функции потерь для случая линейного классификатора

$$L(x, y; w) = \log(1 + \exp(-y \langle w, x \rangle))$$

и упростите итоговое выражение таким образом, чтобы в нём участвовала сигмоидная функция

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

При решении данной задачи вам может понадобиться следующий факт (убедитесь, что он действительно выполняется):

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

Убедимся:

$$\sigma'(z) = \frac{d}{dz}\left(\frac{1}{1 + \exp(-z)}\right) = \frac{\exp(-z)}{(1 + \exp(-z))^2}$$

$$\sigma'(z) = \frac{1}{1 + \exp(-z)} \cdot \left(1 - \frac{1}{1 + \exp(-z)}\right) = \sigma(z)(1 - \sigma(z))$$

Найдём градиент по w:

$$\frac{\partial}{\partial w} L(x, y; w) = \frac{\partial}{\partial w} \log\left(1 + \exp(-y \langle w, x \rangle)\right)$$

$$\frac{\partial}{\partial w} L(x, y; w) = \frac{1}{1 + \exp(-y \langle w, x \rangle)} \cdot \frac{\partial}{\partial w} \exp(-y \langle w, x \rangle).$$

$$\frac{\partial}{\partial w} \exp(-y \langle w, x \rangle) = -y \cdot x \cdot \exp(-y \langle w, x \rangle).$$

Подставляя обратно:

$$\frac{\partial}{\partial w} L(x, y; w) = \frac{\exp(-y \langle w, x \rangle)}{1 + \exp(-y \langle w, x \rangle)} \cdot (-y \cdot x),$$

Упростим:

$$\frac{\partial}{\partial w} L(x, y; w) = -y \cdot x \cdot \frac{\exp(-y \langle w, x \rangle)}{1 + \exp(-y \langle w, x \rangle)}.$$

$$\frac{\exp(-y \langle w, x \rangle)}{1 + \exp(-y \langle w, x \rangle)} = 1 - \sigma(y \langle w, x \rangle).$$

$$\frac{\partial}{\partial w} L(x, y; w) = -y \cdot x \cdot (1 - \sigma(y \langle w, x \rangle)).$$

Также, поскольку $\sigma(-z) = 1 - \sigma(z)$, можно записать:

$$\frac{\partial}{\partial w} L(x, y; w) = -y \cdot x \cdot \sigma(-y \langle w, x \rangle).$$

Задача 8. Ответьте на следующие вопросы:

1. Почему в общем случае распределение $p(y|x)$ для некоторого объекта $x \in \mathbb{X}$ отличается от вырожденного ($p(y|x) \in \{0, 1\}$)?
   Есть несколько возможных причин:

- Вектор признаков x часто не содержит всю информацию, необходимую для однозначного определения метки класса y. Например, при постановке медицинского диагноза набор симптомов может быть характерен для нескольких заболеваний с разной вероятностью. А может быть характерен и для какой-то доли здоровых людей (аналогия для бинарной классификации).
- В реальных данных могут встречаться объекты с идентичными признаковыми описаниями, но разными метками классов, что делает зависимость между x и y неоднозначной.

2. Почему логистическая регрессия позволяет предсказывать корректные вероятности принадлежности объекта классам?
Корректные вероятности будут предсказываться, если loss-функция минимизируется оптимальным константным предсказанием, которое является вероятностью положительного класса на выборке.
Проверим это условие для логистической функции потерь:
Рассмотрим логистическую функцию потерь, где каждому объекту предсказывается одна и та же вероятность $\hat{y} \in (0, 1)$:

$$L(\hat{y}) = -\frac{1}{\ell}\sum_{i=1}^{\ell}\left[y_i \log(\hat{y}) + (1 - y_i)\log(1 - \hat{y})\right].$$

Найдём производную по $\hat{y}$:

$$\frac{dL}{d\hat{y}} = -\frac{1}{\ell}\sum_{i=1}^{\ell}\left(\frac{y_i}{\hat{y}} - \frac{1 - y_i}{1 - \hat{y}}\right).$$

Вынесем средние значения:

$$\frac{dL}{d\hat{y}} = -\left(\frac{1}{\hat{y}} \cdot \frac{1}{\ell}\sum_{i=1}^{\ell} y_i - \frac{1}{1 - \hat{y}} \cdot \frac{1}{\ell}\sum_{i=1}^{\ell}(1 - y_i)\right).$$

Средние значения:

$$\bar{y} = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i, \quad \text{тогда} \quad 1 - \bar{y} = \frac{1}{\ell}\sum_{i=1}^{\ell}(1 - y_i).$$

Подставим:

$$\frac{dL}{d\hat{y}} = -\left(\frac{\bar{y}}{\hat{y}} - \frac{1 - \bar{y}}{1 - \hat{y}}\right).$$

Найдем оптимум:

$$\frac{\bar{y}}{\hat{y}} = \frac{1 - \bar{y}}{1 - \hat{y}}.$$
$$\bar{y}(1 - \hat{y}) = (1 - \bar{y})\hat{y}.$$
$$\bar{y} - \bar{y}\hat{y} = \hat{y} - \bar{y}\hat{y}.$$
$$\bar{y} = \hat{y}.$$

Вывод: оптимальное значение $\hat{y}$, минимизирующее логистическую функцию потерь при константном прогнозе, равно среднему по меткам y_i, что равно доле положительных примеров:

$$\boxed{\hat{y}^* = \bar{y} = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i}.$$

3. Рассмотрим оптимизационную задачу из варианта SVM для линейно разделимых выборок. Всегда ли в обучающей выборке существует объект x_i, для которого выполнено $\langle w, x_i \rangle + b = 1$? Почему?

   Запишем оптимизационную задачу для SVM в случае линейно-разделимой выборки:

$$\begin{cases} ||w||_2^2 \to \min_w \\ y_i(\langle w, x_i \rangle + b) \geq 0 \\ |\langle w, x_i \rangle + b| \geq 1 \end{cases}$$

   Если существует решение задачи, где:

$$\forall x_i \in X : |\langle w, x_i \rangle + b| > 1$$

   то значит, можно подобрать такое $\alpha > 0(1)$, разделить на него w и b, и получить меньшую $||w||_2^2$, а все остальные условия будут продолжать выполняться.
   Иначе: если все условия выполнены, но не существует объекта, для которого выполнено $\langle w, x_i \rangle + b = 1$, значит, $||w||_2^2$ еще имеет пространство для минимизации.

4. С какой целью в постановке оптимизационной задачи SVM для линейно неразделимых выборок вводятся переменные ξ_i, $i = \overline{1, \ell}$?

   ξ_i вводятся для того, чтобы смягчить условия оптимизации (в сравнении с линейно-разделимой выборкой), так как в случае линейно неразделимой выборки такая система не имела бы решения, т.к. невозможно провести гиперплоскость, все отступы объектов от которой были бы положительными (т.е. все объекты были бы классифицированы без ошибок).
   В таком случае решают задачу условной оптимизации следующего вида:

$$\begin{cases} ||w||_2^2 + C\frac{1}{l}\sum_{i=1}^{l} \xi_i \to \min_{w,b,\xi_i} \\ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

   Таким образом, мы позволяем объектам находиться с ошибочной стороны от гиперплоскости (ξ_i таких объектов > 0) и одновременно добавляем в наш функционал среднее ξ_i, чтобы избежать тривиального решения $\xi_i \to \infty, ||w||_2^2 = 0$