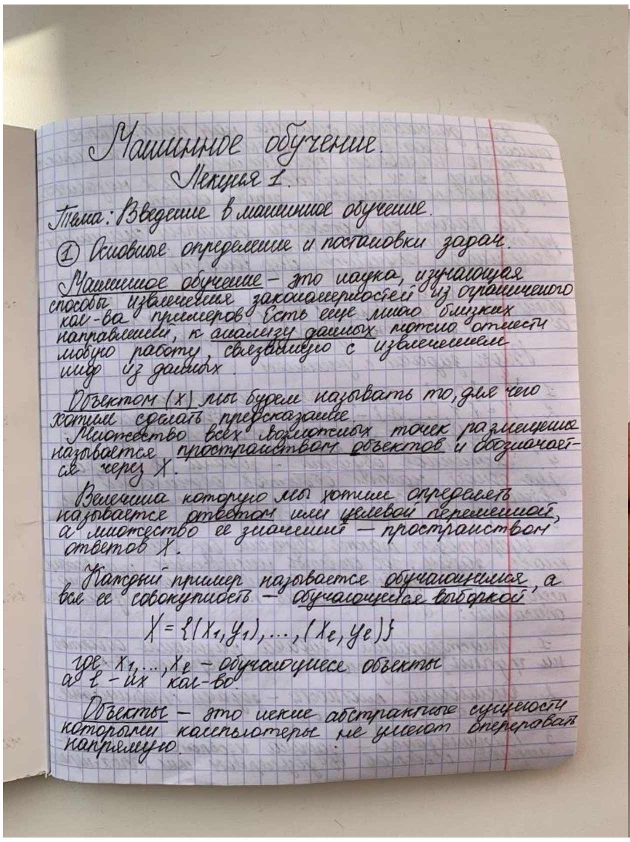


Машинное обучение.

Лекция 1.

Тема: Введение в машинное обучение.

① Основные определения и постановки задач.

Машинное обучение – это наука, изучающая способы извлечения закономерностей из ограниченного кол-ва примеров. Есть еще много близких направлений, к анализу данных можно отнести любую работу, связанную с извлечением инф из данных.

Объектом (x) мы будем называть то, для чего хотим сделать предсказание.
Множество всех возможных точек размещения называется пространством объектов и обозначается через X.

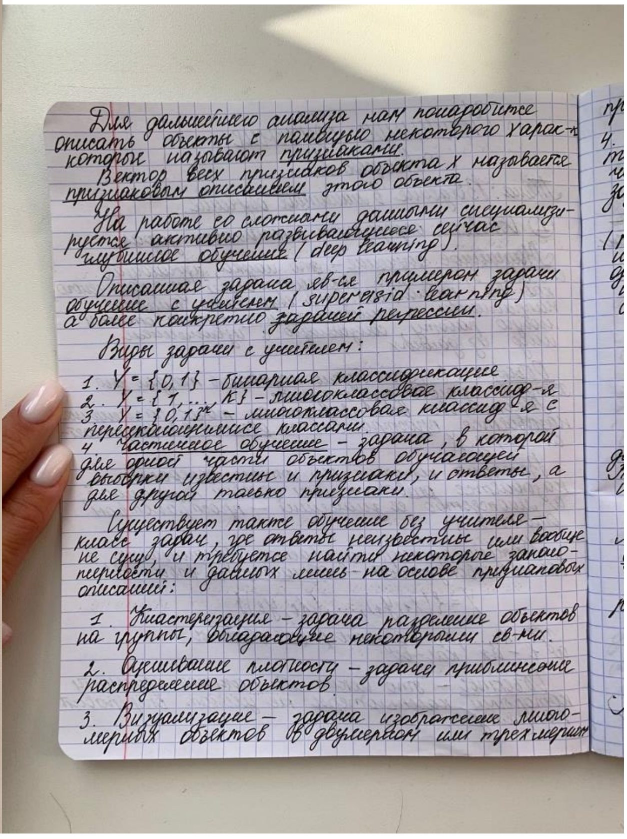
Величина которую мы хотим определять называется ответом или целевой переменной, а множество ее значений – пространством ответов X.

Каждый пример называется обучающимся, а вся ее совокупность – обучающейся выборкой

$$X = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$$

где $x_1, \dots, x_\ell$ – обучающиеся объекты а ℓ – их кол-во.

Объекты – это некие абстрактные сущности которыми компьютеры не умеют оперировать напрямую.



Для дальнейшего анализа нам понадобится описать объекты с помощью некоторого характ-к, которые называют признаками.
Вектор всех признаков объекта x называется признаковым описанием этого объекта.

На работе со сложными данными специализируется активно развивающееся сейчас глубинное обучение (deep learning).

Описанная задача яв-се примером задачи обучение с учителем (supervisid learning) а более конкретно задачей регрессии.

Виды задачи с учителем:

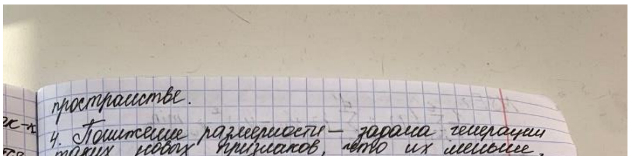
1. $Y = \{0, 1\}$ – бинарная классификация
2. $Y = \{1, \dots, K\}$ – многоклассовая классиф-я
3. $Y = \{0, 1\}^K$ – многоклассовая классиф-я с пересекающимися классами.
4. Частичное обучение – задача, в которой для одной части объектов обучающей выборки известны и признаки, и ответы, а для другой только признаки.

Существует также обучение без учителя – класс задач, где ответы неизвестны или вообще не суще, и требуется найти некоторые закономерности в данных лишь на основе признаковых описаний:

1. Кластеризация – задача разделения объектов на группы, обладающие некоторыми св-ми.

2. Оценивание плотности – задача приближения распределения объектов.

3. Визуализация – задача изображения многомерных объектов в двумерном или трехмерном



пространстве.

4. Понижение размерности – задача генерации таких новых признаков, что их меньше, чем исходных, но при этом с их помощью задача решается хуже.

Сложные постановки – обучение с подкреплением (reinforcement learning) где алгоритм на каждом шаге наблюдает какую-то ситуацию, выбирает одно из доступных ему действий, получает некоторую награду и корректирует свою стратегию.

Объекты-признаки

$$X \in R^{\ell \times d}$$

ℓ - число объектов
d - число признаков.

Построение функции $a: X \to Y$, которая для любого объекта будет предсказывать ответ. Такая функция называется алгоритмом или моделью.

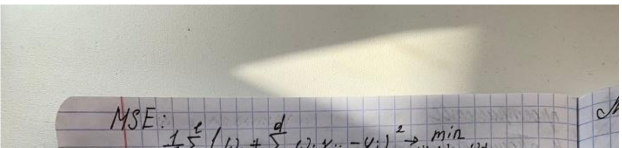
Если функционал устроен так, что его следует минимизировать, то логично называть его функционал ошибки.

Крайне популярным функционалом в задаче регрессии яв-се среднеквадратичная ошибка:

$$Q(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}(a(X_i)-(y_i))^2$$

Линейные модели

$$A = \{a(x) = w_0 + w_1x_1 + \dots + w_dx_d \mid w_0, w_1, \dots, w_d \in R\}$$



MSE:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}\left(w_0 + \sum_{j=1}^{d} w_jx_{ij} - y_i\right)^2 \to \min_{w_0, w_1 \dots w_d}$$

Процесс поиска оптимального алгоритма называется обучением.

Контроль сложности семейства алгоритмов – чем меньше у нас данных для обучения, тем более простые семейства следует выбирать.

Этапы решения задачи машинного обучения:

1. Постановка задачи.
2. Выделение признаков.
3. Формирование выборки.
4. Выбор функционала ошибки.
5. Предобработка данных
6. Построение модели.
7. Оценивание качества модели.

Лекция 2.

Тема: Линейная регрессия.

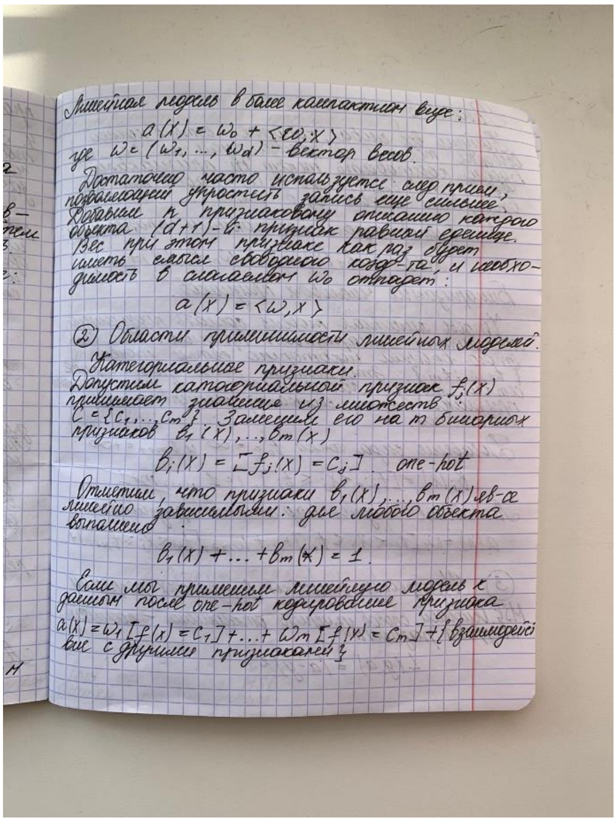
① Линейные модели.

Модели сводятся к суммированию значений признаков с некоторыми весами:

$$a(x) = w_0 + \sum_{j=1}^{d} w_jx_j$$

Параметрами модели яв-се веса или коэффициенты w_j.

Вес w_0 также называются свободным коэф-н или сдвигом (bias).



Линейная модель в более компактном виде:

$$a(X) = w_0 + \langle w, x \rangle$$

где $w = (w_1, \dots, w_d)$ – вектор весов.

Достаточно часто используется след прием, позволяющий упростить запись еще сильнее. Добавим к признаковому описанию каждого объекта $(d+1)$-й признак равный единице. Вес при этом признаке как раз будет иметь смысл свободного коэф-та, и необходимость в слагаемом w_0 отпадет:

$$a(X) = \langle w, x \rangle$$

② Области применимости линейных моделей.

Категориальные признаки.

Допустим категориальный признак $f_j(X)$ принимает значение из множеств: $C = \{c_1, \dots, c_m\}$. Заменим его на m бинарных признаков $b_1(X), \dots, b_m(X)$

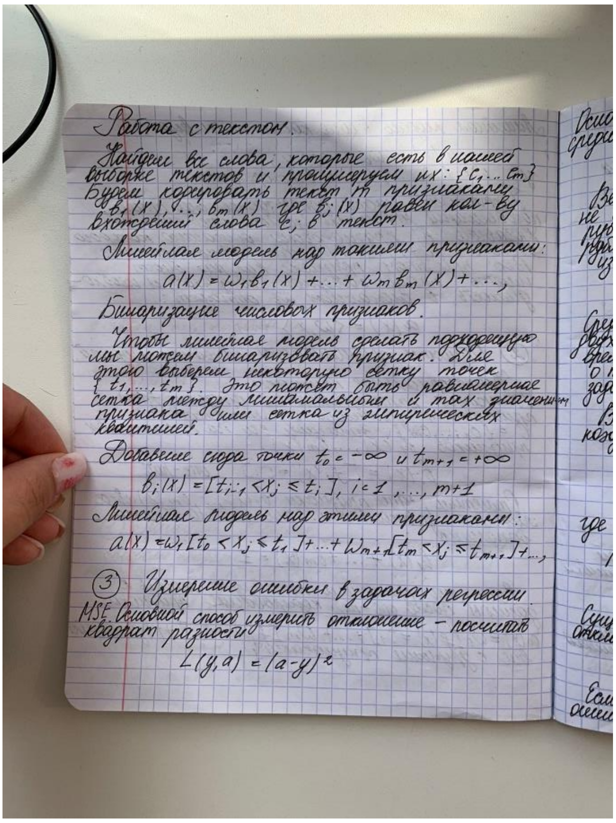
$$b_i(X) = [f_j(X) = c_j]. \quad \text{one-hot}$$

Отметим, что признаки $b_1(X), \dots, b_m(X)$ яв-ся линейно зависимыми: для любого объекта выполнено

$$b_1(X) + \dots + b_m(X) = 1.$$

Если мы применим линейную модель к данным после one-hot кодирования признака

$$a(X) = w_1[f(x) = c_1] + \dots + w_m[f(X) = c_m] + \{\text{взаимодействие с другими признаками}\}$$



Работа с текстом.

Найдем все слова, которые есть в нашей выборке текстов и пронумеруем их: $\{c_1, \dots c_m\}$

Будем кодировать текст m признаками $b_1(X), \dots, b_m(X)$ где $b_i(X)$ равен кол-ву вхождений слова c_i в текст.

Линейная модель над такими признаками:

$$a(X) = w_1 b_1(X) + \dots + w_m b_m(X) + \dots,$$

Бинаризация числовых признаков.

Чтобы линейная модель сделать подходящую мы можем бинаризовать признак. Для этого выберем некоторую сетку точек $\{t_1, \dots, t_m\}$. Это может быть равномерная сетка между минимальным и max значением признака или сетка из эмпирических квантилей.

Добавим сюда точки $t_0 = -\infty$ и $t_{m+1} = +\infty$

$$b_i(X) = [t_{i-1} < X_j \leq t_i], \; i = 1, \dots, m+1$$

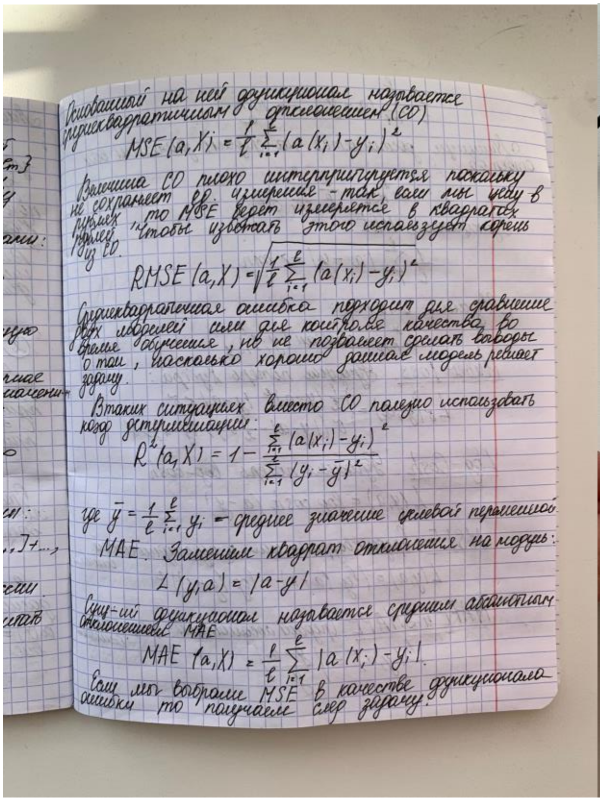
Линейная модель над этими признаками:

$$a(X) = w_1[t_0 < X_j \leq t_1] + \dots + w_{m+1}[t_m < X_j \leq t_{m+1}] + \dots,$$

③ Измерение ошибки в задачах регрессии.

MSE. Основной способ измерить отклонение – посчитать квадрат разности

$$L(y,a) = (a-y)^2$$



Основанный на ней функционал называется среднеквадратичным отклонением (СО)

$$MSE(a,X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2.$$

Величина СО плохо интерпретируется поскольку не сохраняет ед. измерения – так, если мы цену в рублях, то MSE будет измеряться в квадратах рублей. Чтобы избежать этого используют корень из СО.

$$RMSE(a,X) = \sqrt{\frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2}$$

Среднеквадратичная ошибка подходит для сравнения двух моделей или для контроля качества во время обучения, но не позволяет сделать выводы о том, насколько хорошо данная модель решает задачу.

В таких ситуациях вместо СО полезно использовать коэф детерминации:

$$R^2(a,X) = 1 - \frac{\sum_{i=1}^{l}(a(x_i) - y_i)^2}{\sum_{i=1}^{l}(y_i - \bar{y})^2}$$

где $\bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$ – среднее значение целевой переменной

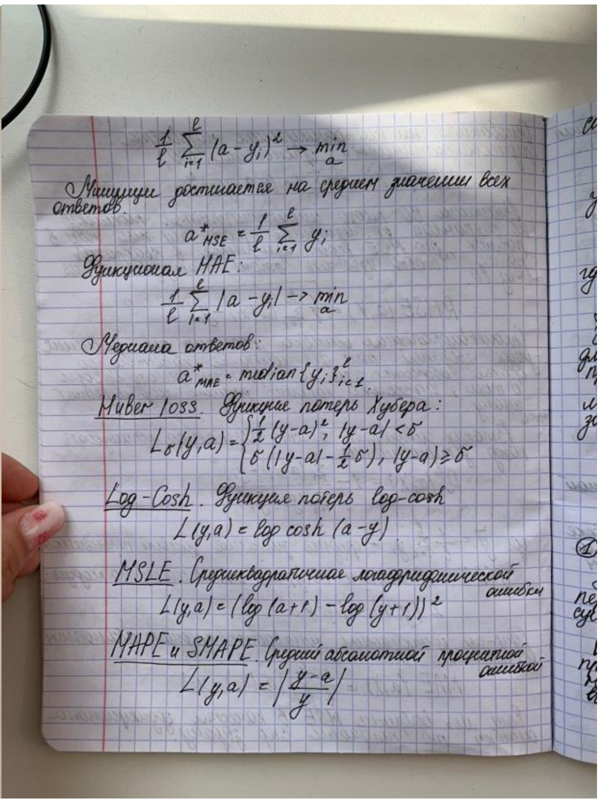
MAE. Заменим квадрат отклонения на модуль:

$$L(y,a) = |a-y|$$

Соотв-ий функционал называется средним абсолютным отклонением MAE

$$MAE(a,X) = \frac{l}{l}\sum_{i=1}^{l}|a(x_i) - y_i|.$$

Если мы выбрали MSE в качестве функционала ошибки то получаем след задачу:



$$\frac{1}{l}\sum_{i=1}^{l}(a - y_i)^2 \to \min_a$$

Минимум достигается на среднем значении всех ответов.

$$a^*_{MSE} = \frac{1}{l}\sum_{i=1}^{l} y_i$$

Функционал MAE:

$$\frac{1}{l}\sum_{i=1}^{l}|a - y_i| \to \min_a$$

Медиана ответов:

$$a^*_{MAE} = median\{y_i\}_{i=1}^{l}.$$

Huber loss. Функция потерь Хубера:

$$L_\delta(y,a) = \begin{cases} \frac{1}{2}(y-a)^2, & |y-a| < \delta \\ \delta(|y-a| - \frac{1}{2}\delta), & |y-a| \geq \delta \end{cases}$$

Log-Cosh. Функция потерь log-cosh

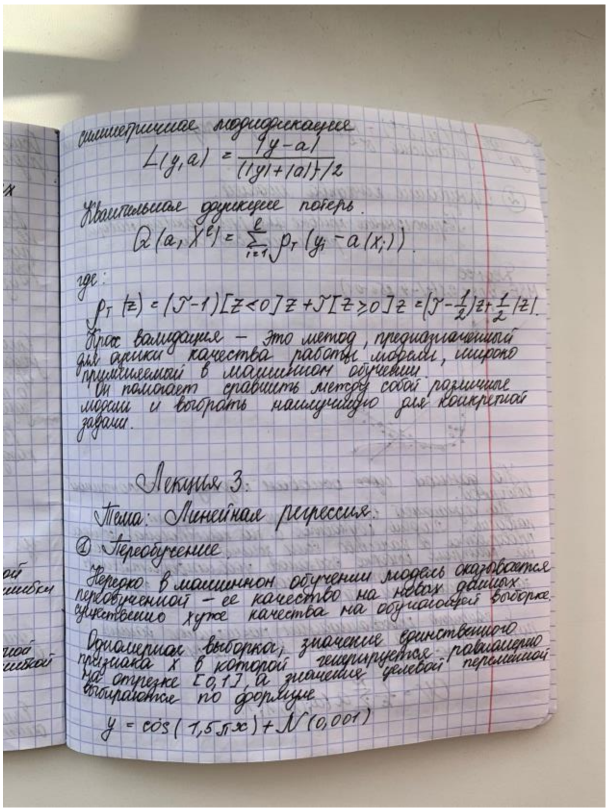
$$L(y,a) = \log\cosh(a-y)$$

MSLE. Среднеквадратичная логарифмической ошибки

$$L(y,a) = (\log(a+1) - \log(y+1))^2$$

MAPE и SMAPE. Средней абсолютной процентной ошибкой

$$L(y,a) = \left|\frac{y-a}{y}\right|$$



симметричная модификация

$$L(y,a) = \frac{|y-a|}{(|y|+|a|)/2}$$

Квантильная функция потерь.

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} \rho_\tau (y_i - a(x_i)).$$

где:

$$\rho_\tau(z) = (\tau-1)[z<0]z + \tau[z \geq 0]z = (\tau - \tfrac{1}{2})z + \tfrac{1}{2}|z|.$$

Кросс валидация – это метод, предназначенный для оценки качества работы модели, широко применяемой в машинном обучении.

Он помогает сравнить между собой различные модели и выбрать наилучшую для конкретной задачи.

Лекция 3.

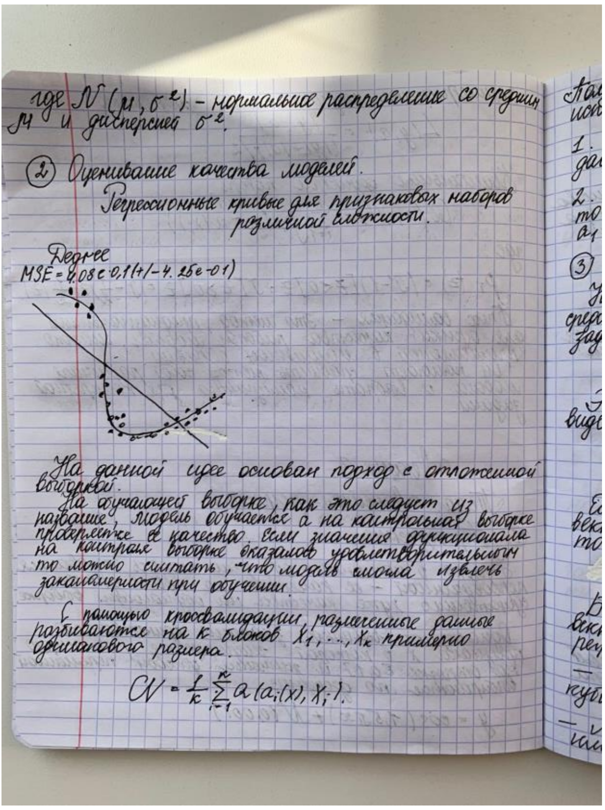
Тема: Линейная регрессия.

① Переобучение.

Нередко в машинном обучении модель оказывается переобученной – ее качество на новых данных существенно хуже качества на обучающей выборке.

Одномерная выборка, значение единственного признака x в которой генерируется равномерно на отрезке $[0,1]$, а значения целевой переменной выбираются по формуле

$$y = \cos(1{,}5\pi x) + N(0{,}001)$$



где $N(\mu, \sigma^2)$ – нормальное распределение со средним μ и дисперсией σ^2.

② Оценивание качества моделей.

Регрессионные кривые для признаковых наборов различной сложности.

Degree

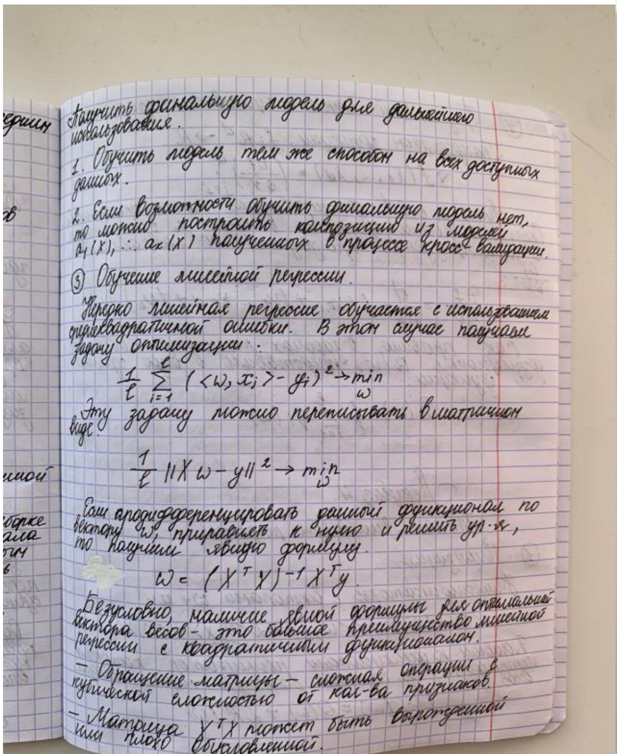
MSE = 4.08e-0.1(+/-4.25e-01)

На данной идее основан подход с отложенной выборкой.

На обучающей выборке, как это следует из названия, модель обучается а на контрольной выборке проверяется ее качество. Если значения функционала на контроле выборке оказалось удовлетворительным то можно считать, что модель смогла извлечь закономерности при обучении.

С помощью кроссвалидации, размеченные данные разбиваются на k блоков $X_1, \ldots, X_k$ примерно одинакового размера.

$$CV = \frac{1}{k} \sum_{i=1}^{k} Q(a_i(X), X_i).$$



Получить финальную модель для дальнейшего использования.

1. Обучить модель тем же способом на всех доступных данных.

2. Если возможности обучить финальную модель нет, то можно построить композицию из моделей $a_1(X), \ldots a_k(X)$ полученных в процессе кросс-валидации.

③ Обучение линейной регрессии.

Нередко линейная регрессия обучается с использованием среднеквадратичной ошибки. В этом случае получаем задачу оптимизации:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \to \min_w$$

Эту задачу можно переписывать в матричном виде:

$$\frac{1}{\ell} \|Xw - y\|^2 \to \min_w$$

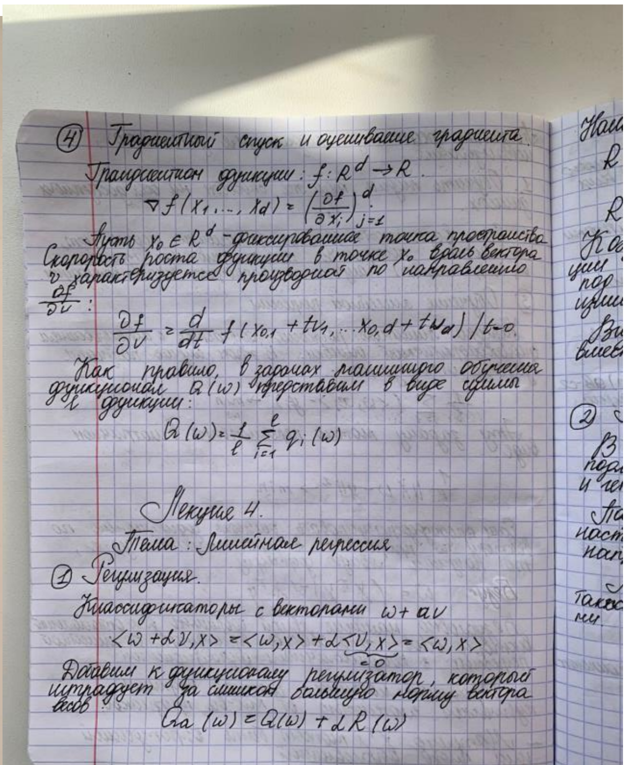
Если продифференцировать данный функционал по вектору w, приравнять к нулю и решить ур-я, то получим явную формулу.

$$w = (X^T X)^{-1} X^T y.$$

Безусловно, наличие явной формулы для оптимального вектора весов – это большое преимущество линейной регрессии с квадратичным функционалом.

– Обращение матрицы – сложная операция с кубической сложностью от кол-ва признаков.

– Матрица $X^T X$ может быть вырожденной или плохо обусловленной.



④ Градиентный спуск и оценивание градиента.

Градиентом функции: $f: R^d \to R$.

$$\nabla f(x_1, \ldots, x_d) = \left(\frac{\partial f}{\partial x_i}\right)_{j=1}^{d}.$$

Пусть $x_0 \in R^d$ – фиксированная точка пространства. Скорость роста функции в точке x_0 вдоль вектора v характеризуется производная по направлению $\frac{\partial f}{\partial v}$:

$$\frac{\partial f}{\partial v} = \frac{d}{dt} f(x_{0,1} + tv_1, \ldots x_{0,d} + tv_d)|_{t=0}.$$

Как правило, в задачах машинного обучения функционал $Q(w)$ представим в виде суммы ℓ функций:

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} q_i(w)$$

Лекция 4.

Тема: Линейная регрессия

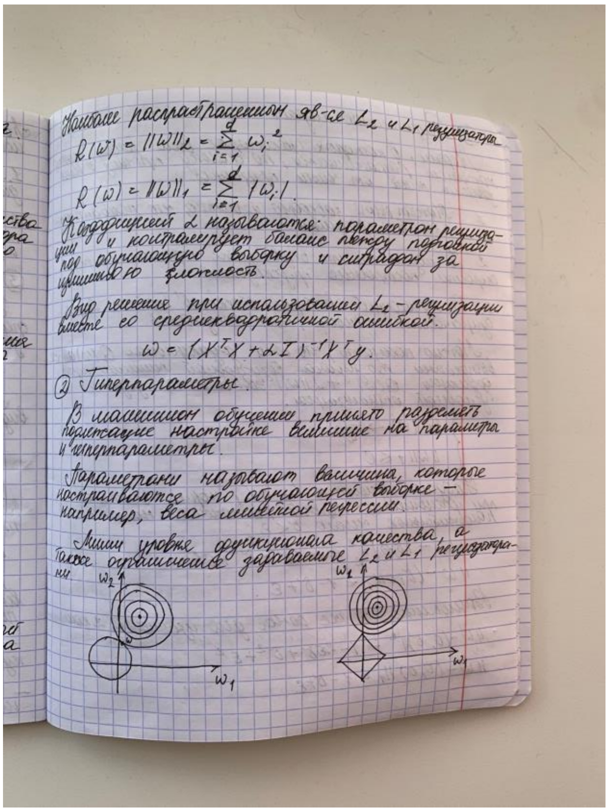
① Регуляризация.

Классификаторы с векторами $w + \alpha v$

$$\langle w + \alpha v, x \rangle = \langle w, x \rangle + \alpha \underbrace{\langle v, x \rangle}_{=0} = \langle w, x \rangle$$

Добавим к функционалу регуляризатор, который штрафует за слишком большую норму вектора весов:

$$Q_\alpha(w) = Q(w) + \alpha R(w)$$



Наиболее распространенным яв-се L_2 и L_1 регуляризатора

$$R(w) = \|w\|_2 = \sum_{i=1}^{d} w_i^2$$

$$R(w) = \|w\|_1 = \sum_{i=1}^{d} |w_i|.$$

Коэффициент α называются параметром регуляризации и контролирует баланс между подгонкой под обучающую выборку и штрафом за излишнюю сложность.

Вид решения при использовании L_2-регуляризации вместе со среднеквадратичной ошибкой.

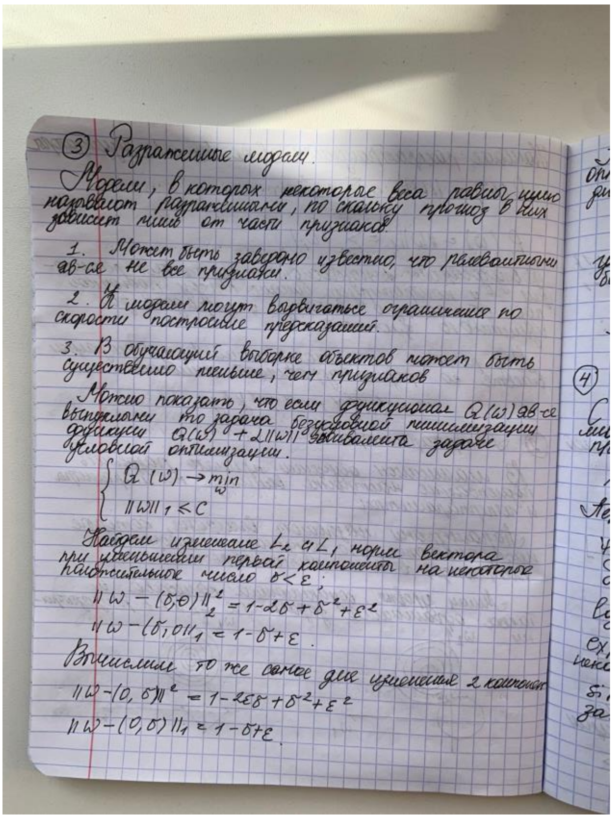
$$w = (X^T X + \alpha I)^{-1} X^T y.$$

② Гиперпараметры.

В машинном обучении принято разделять подлежащие настройке величины на параметры и гиперпараметры.

Параметрами называют величина, которые настраиваются по обучающей выборке — например, веса линейной регрессии.

Линии уровне функционала качества, а также ограничение задаваемые L_2 и L_1 регуляризаторами



③ Разреженные модели.

Модели, в которых некоторые веса равны нулю называют разреженными, поскольку прогноз в них зависит лишь от части признаков.

1. Может быть заведомо известно, что релевантными яв-се не все признаки.

2. К модели могут выдвигаться ограничения по скорости построения предсказаний.

3. В обучающей выборке объектов может быть существенно меньше, чем признаков

Можно показать, что если функционал $Q(w)$ яв-се выпуклым то задача безусловной минимизации функции $Q(w) + \alpha\|w\|$ эквивалента задаче условной оптимизации.

$$\begin{cases} Q(w) \to \min\limits_{w} \\ \|w\|_1 \leq C \end{cases}$$

Найдем изменение L_2 и L_1 норм вектора при уменьшении первой компоненты на некоторое положительное число $\delta < \varepsilon$;

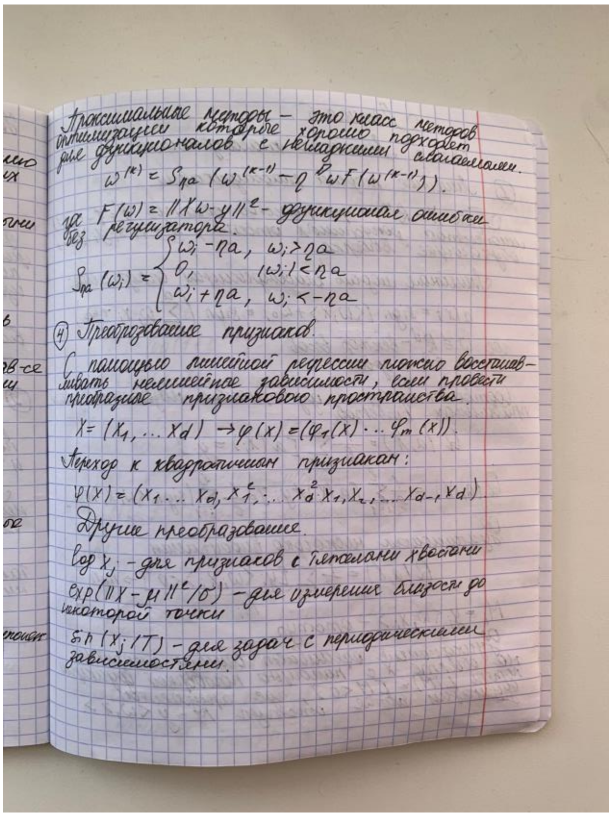
$$\|w - (\delta, 0)\|_2^2 = 1 - 2\delta + \delta^2 + \varepsilon^2$$

$$\|w - (\delta, 0\|_1 = 1 - \delta + \varepsilon.$$

Вычислим то же самое для изменения 2 компонент

$$\|w - (0, \delta)\|^2 = 1 - 2\varepsilon\delta + \delta^2 + \varepsilon^2$$

$$\|w - (0, \delta)\|_1 = 1 - \delta + \varepsilon.$$



Проксимальные методы — это класс методов оптимизации которые хорошо подходят для функционалов с негладкими слагаемыми.

$$w^{(k)} = S_{\eta a}(w^{(k-1)} - \eta \nabla_w F(w^{(k-1)})).$$

где $F(w) = \|Xw - y\|^2$ — функционал ошибки без регуляризатора.

$$S_{\eta a}(w_i) = \begin{cases} w_i - \eta a, & w_i > \eta a \\ 0, & |w_i| < \eta a \\ w_i + \eta a, & w_i < -\eta a \end{cases}$$

④ Преобразование признаков.

С помощью линейной регрессии можно восстанавливать нелинейные зависимости, если провести преобразование признакового пространства.

$$X = (x_1, \dots x_d) \to \varphi(x) = (\varphi_1(x) \cdots \varphi_m(x)).$$

Переход к квадратичным признакам:

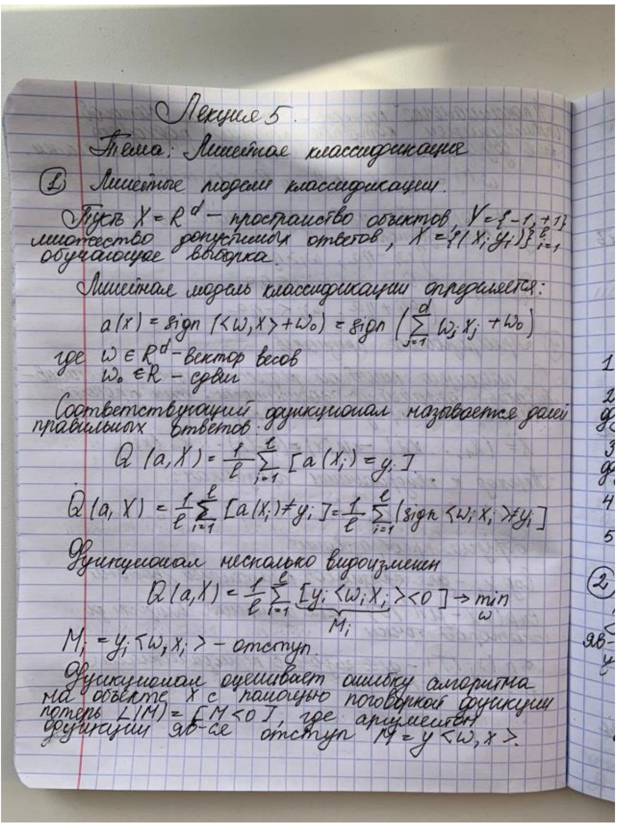
$$\varphi(x) = (x_1 \dots x_d, x_1^2, \dots x_d^2 x_1, x_2, \dots x_{d-1} x_d).$$

Другие преобразования.

$\log x_j$ — для признаков с тяжелыми хвостами

$\exp(\|x - \mu\|^2 / \sigma)$ — для измерения близости до некоторой точки

$\sin(x_j / T)$ — для задач с периодическими зависимостями.



Лекция 5.

Тема: Линейная классификация

① Линейные модели классификации.

Пусть $X = R^d$ — пространство объектов, $Y = \{-1, +1\}$ множество допустимых ответов, $X = \{(x_i, y_i)\}_{i=1}^{\ell}$ обучающая выборка.

Линейная модель классификации определяется:

$$a(x) = \operatorname{sign}(\langle w, x\rangle + w_0) = \operatorname{sign}\left(\sum_{j=1}^{d} w_j x_j + w_0\right)$$

где $w \in R^d$ — вектор весов
$w_0 \in R$ — сдвиг

Соответствующий функционал называется долей правильных ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i].$$

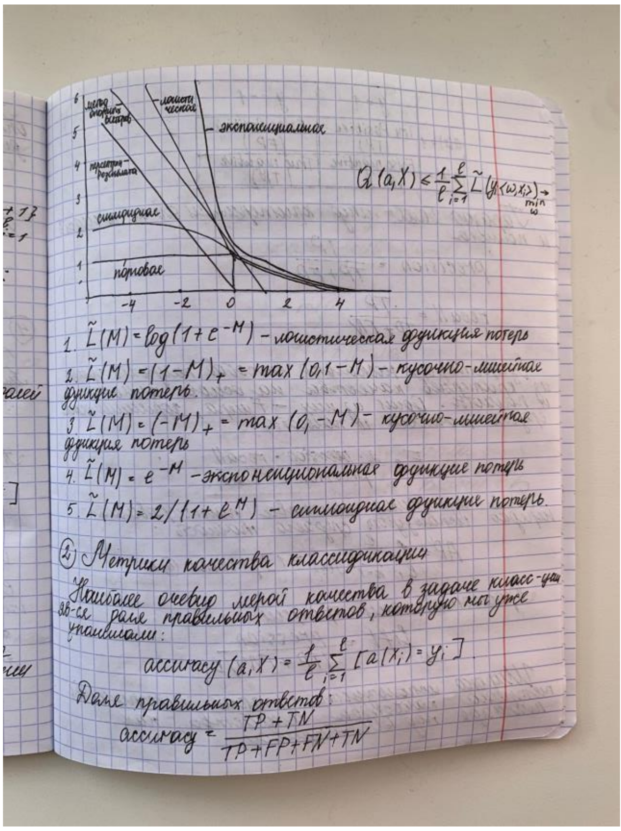
$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i] = \frac{1}{\ell} \sum_{i=1}^{\ell} [\operatorname{sign}\langle w_i x_i\rangle \neq y_i]$$

Функционал несколько видоизменен

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [\underbrace{y_i \langle w_i x_i}_{M_i}\rangle < 0] \to \min_{w}$$

$M_i = y_i \langle w, x_i\rangle$ — отступ.

Функционал оценивает ошибку алгоритма на объекте x с помощью пороговой функции потерь $L(M) = [M < 0]$, где аргументом функции яв-се отступ $M = y\langle w, x\rangle$.


6
5
4
3
2
1
-4
-2
0
2
4
метод опорных векторов
логистическая
экспоненциальная
персептрон-Розенблата
сигмоидная
пороговая

$$Q(a,X) \leq \frac{1}{\ell}\sum_{i=1}^{\ell}\tilde{L}(y_i\langle w,x_i\rangle) \to \min_{w}$$

1. $\tilde{L}(M)=\log(1+e^{-M})$ – логистическая функция потерь

2. $\tilde{L}(M)=(1-M)_+ = \max(0,1-M)$ – кусочно-линейная функция потерь.

3. $\tilde{L}(M)=(-M)_+ = \max(0,-M)$ – кусочно-линейная функция потерь

4. $\tilde{L}(M)=e^{-M}$ – экспоненциальная функция потерь

5. $\tilde{L}(M)=2/(1+e^{M})$ – сигмоидная функция потерь.

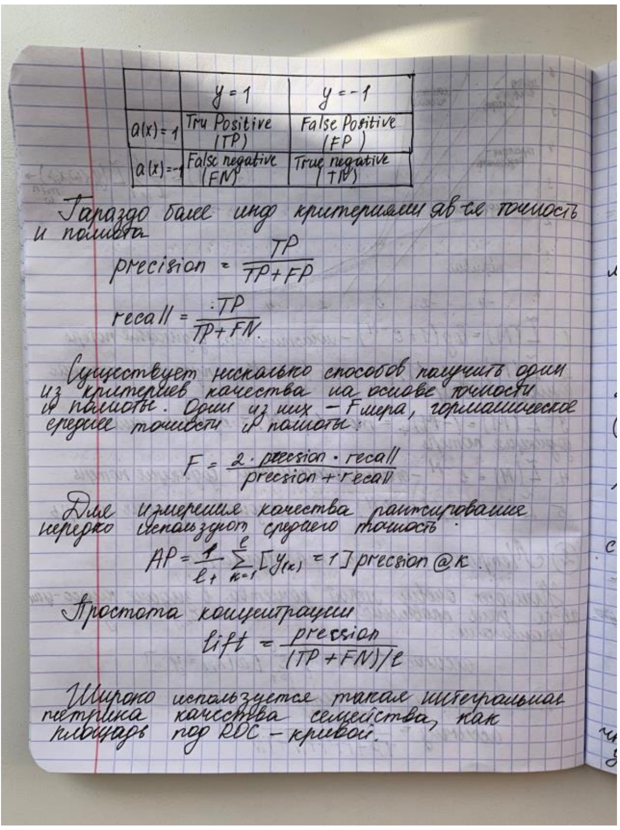
② Метрики качества классификации

Наиболее очевид. мерой качества в задаче класс-ции яв-ся доля правильных ответов, которую мы уже упоминали:

$$accuracy(a,X)=\frac{1}{\ell}\sum_{i=1}^{\ell}[a(x_i)=y_i].$$

Доля правильных ответов:

$$accuracy=\frac{TP+TN}{TP+FP+FN+TN}$$



	$y=1$	$y=-1$
$a(x)=1$	Tru Positive (TP)	False Positive (FP)
$a(x)=-1$	False negative (FN)	True negative (TN)

Гораздо более инф. критериями яв-ся точность и полнота

$$precision=\frac{TP}{TP+FP}$$

$$recall=\frac{TP}{TP+FN}.$$

Существует несколько способов получить один из критериев качества на основе точности и полноты. Один из них – F мера, гармоническое среднее точности и полноты:

$$F=\frac{2\cdot precision\cdot recall}{precision+recall}$$

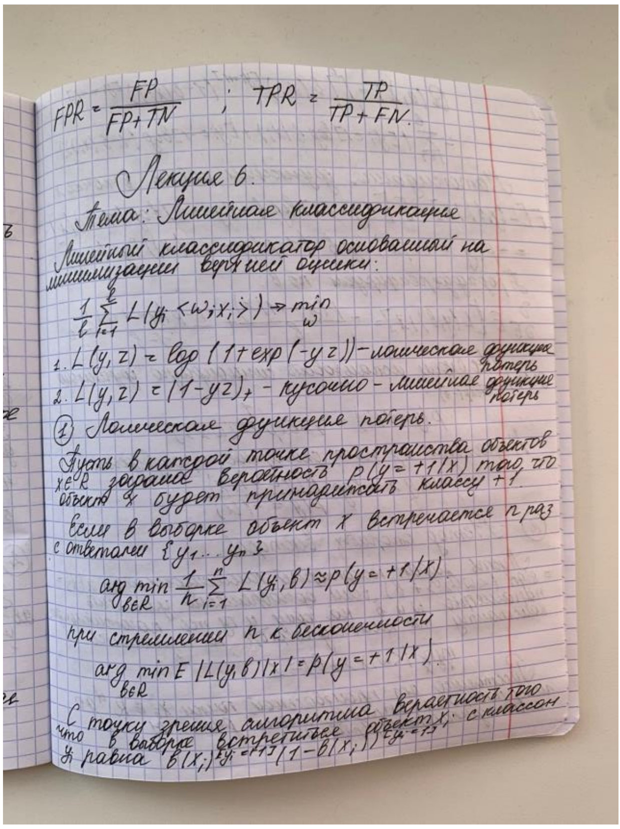
Для измерения качества ранжирования нередко используют среднего точность:

$$AP=\frac{1}{\ell_+}\sum_{k=1}^{\ell}[y_{(k)}=1]\,precision@k.$$

Простота концентрации

$$lift=\frac{precision}{(TP+FN)/\ell}$$

Широко используется такая интегральная метрика качества семейства, как площадь под ROC – кривой.



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

Лекция 6.

Тема: Линейная классификация

Линейный классификатор основанный на минимизации верхней оценки:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}L(y_i\langle w;x_i\rangle)\to\min_{w}$$

1. $L(y,z)=\log(1+\exp(-yz))$ – логическая функция потерь

2. $L(y,z)=(1-yz)_+$ – кусочно-линейная функция потерь

① Логическая функция потерь.

Пусть в каждой точке пространства объектов $x\in R$ задана вероятность $p(y=+1|x)$ того, что объект x будет принадлежать классу +1.

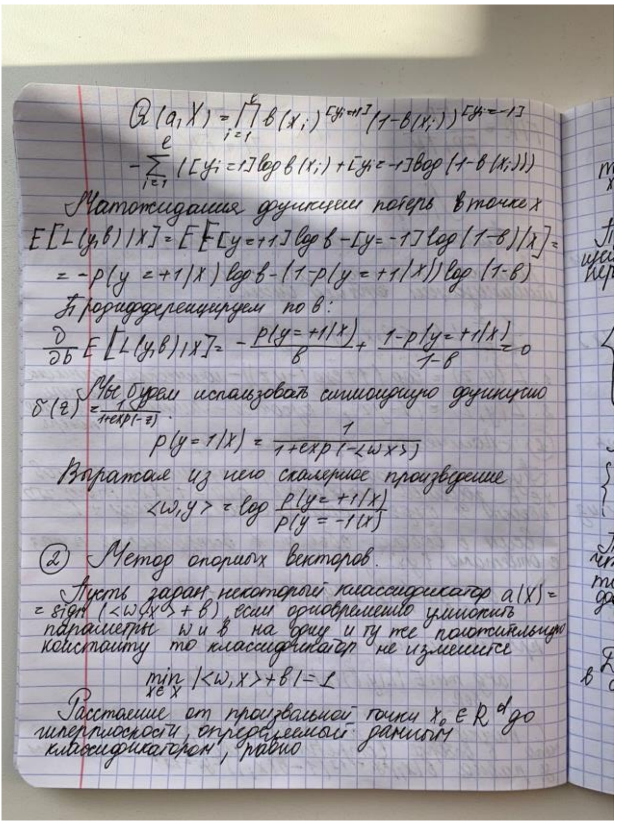
Если в выборке объект x встречается n раз с ответами $\{y_1 \dots y_n\}$

$$\arg\min_{b\in R}\frac{1}{n}\sum_{i=1}^{n}L(y_i,b)\approx p(y=+1|x).$$

при стремлении n к бесконечности

$$\arg\min_{b\in R}E[L(y,b)|x]=p(y=+1|x).$$

С точки зрения алгоритма вероятность того что в выборке встретится объект x_i с классом y_i равна $b(x_i)^{[y_i=+1]}(1-b(x_i))^{[y_i=-1]}$



$$Q(a,X)=\prod_{i=1}^{\ell}b(x_i)^{[y_i=+1]}(1-b(x_i))^{[y_i=-1]}$$

$$-\sum_{i=1}^{\ell}([y_i=1]\log b(x_i)+[y_i=-1]\log(1-b(x_i)))$$

Матожидания функции потерь в точке x

$$E[L(y,b)|x]=E[-[y=+1]\log b-[y=-1]\log(1-b)|x]=$$

$$=-p(y=+1|x)\log b-(1-p(y=+1|x))\log(1-b)$$

Продифференцируем по b:

$$\frac{\partial}{\partial b}E[L(y,b)|x]=-\frac{p(y=+1|x)}{b}+\frac{1-p(y=+1|x)}{1-b}=0$$

Мы будем использовать сигмоидную функцию $\sigma(z)=\frac{1}{1+\exp(-z)}$.

$$p(y=1|x)=\frac{1}{1+\exp(-\langle w x\rangle)}$$

Выражая из него скалярное произведение

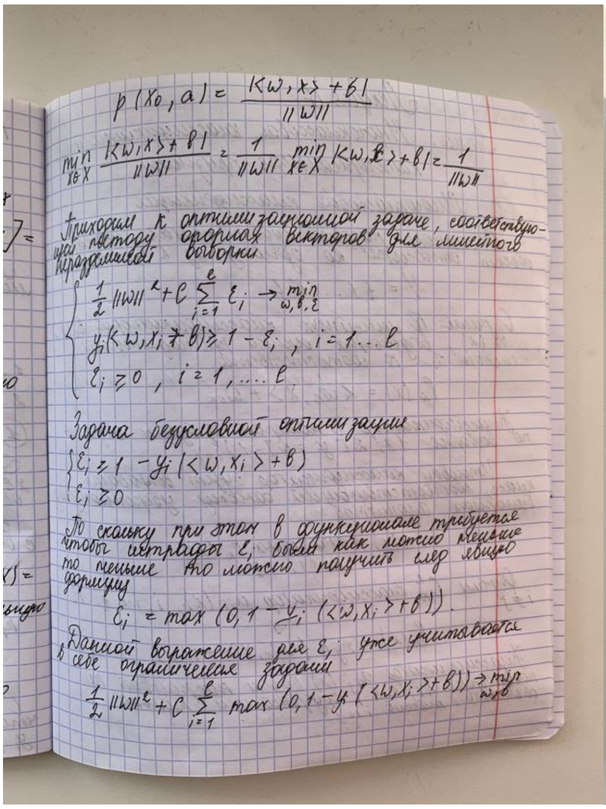
$$\langle w,y\rangle=\log\frac{p(y=+1|x)}{p(y=-1|x)}$$

② Метод опорных векторов.

Пусть задан некоторый классификатор $a(x)=\operatorname{sign}(\langle w,x\rangle+b)$, если одновременно умножить параметры w и b на одну и ту же положительную константу то классификатор не изменится

$$\min_{x\in X}|\langle w,x\rangle+b|=1$$

Расстояние от произвольной точки $x_0\in R^d$ до гиперплоскости, определяемой данным классификатором, равно



$$p(x_0, a) = \frac{|\langle w, x\rangle + b|}{\|w\|}$$

$$\min_{x\in X} \frac{|\langle w, x\rangle + b|}{\|w\|} = \frac{1}{\|w\|} \min_{x\in X} |\langle w, x\rangle + b| = \frac{1}{\|w\|}$$

Приходим к оптимизационной задаче, соответствующей методу опорных векторов для линейно неразделимой выборки

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

Задача безусловной оптимизации

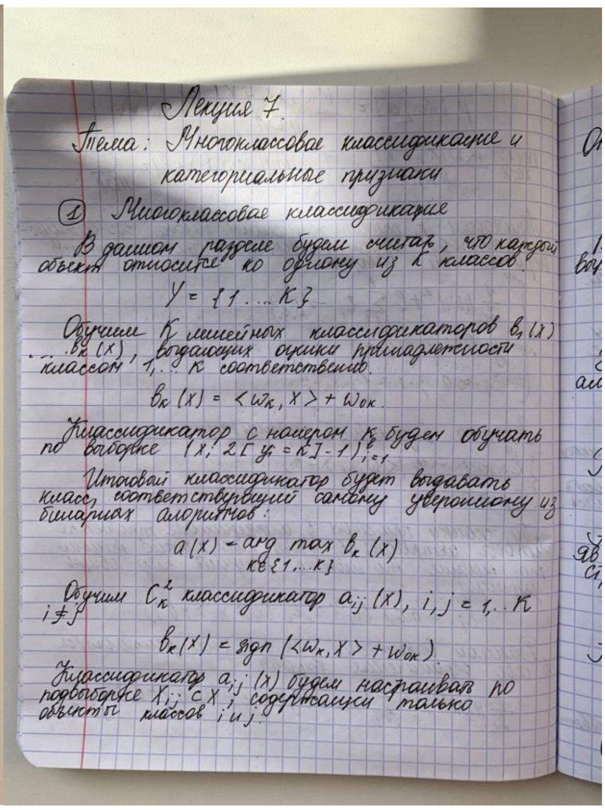
$$\begin{cases} \varepsilon_i \geq 1 - y_i(\langle w, x_i\rangle + b) \\ \varepsilon_i \geq 0 \end{cases}$$

Поскольку при этом в функционале требуется чтобы штрафы ε_i были как можно меньше то можно получить след явную формулу

$$\varepsilon_i = \max(0, 1 - y_i(\langle w, x_i\rangle + b)).$$

Данное выражение для ε_i уже учитывается в себе ограничения задачи

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} \max(0, 1 - y_i(\langle w, x_i\rangle + b)) \to \min_{w,b}$$



Лекция 7.

Тема: Многоклассовая классификация и категориальные признаки

① Многоклассовая классификация

В данном разделе будем считать, что каждый объект относится ко одному из K классов.

$$Y = \{1 \ldots K\}$$

Обучим K линейных классификаторов $b_1(x) \ldots b_K(x)$, выдающих оценки принадлежности классам 1, .. K соответственно.

$$b_k(x) = \langle w_k, x\rangle + w_{0k}.$$

Классификатор с номером k будем обучать по выборке $(x_i, 2[y_i = k] - 1)_{i=1}^{\ell}$

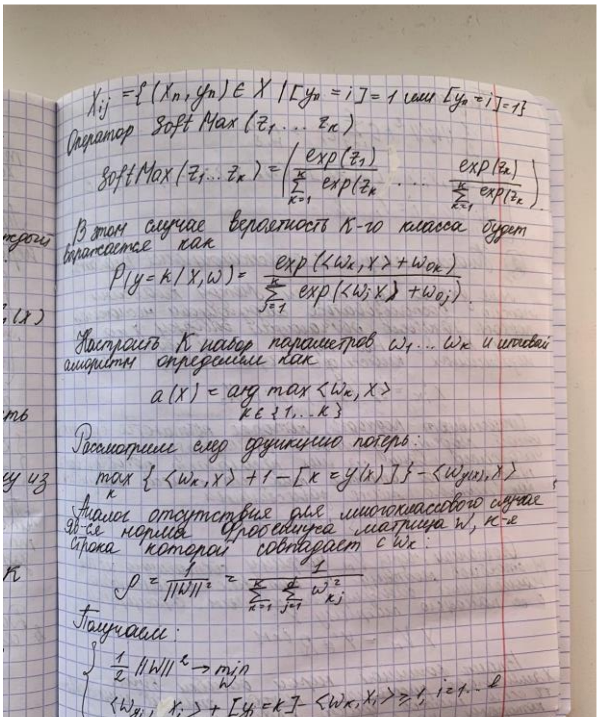
Итоговый классификатор будет выдавать класс, соответствующий самому уверенному из бинарных алгоритмов:

$$a(x) = \arg\max_{k\in\{1,\ldots K\}} b_k(x)$$

Обучим C_K^2 классификатор $a_{ij}(x)$, $i, j = 1, .. K$ $i \neq j$

$$b_k(x) = \text{sign}(\langle w_k, x\rangle + w_{0k}).$$

Классификатор $a_{ij}(x)$ будем настраивать по подвыборке $X_{ij} \subset X$, содержащей только объекты классов i и j.



$$X_{ij} = \{(x_n, y_n) \in X \mid [y_n = i] = 1 \text{ или } [y_n = i] = 1\}$$

Оператор $\text{SoftMax}(z_1 \ldots z_K)$

$$\text{SoftMax}(z_1 \ldots z_K) = \left(\frac{\exp(z_1)}{\sum_{k=1}^{K}\exp(z_k)} \ldots \frac{\exp(z_K)}{\sum_{k=1}^{K}\exp(z_k)}\right).$$

В этом случае вероятность k-го класса будет выражаться как

$$P(y = k / x, w) = \frac{\exp(\langle w_k, x\rangle + w_{0k})}{\sum_{j=1}^{K}\exp(\langle w_j, x\rangle + w_{0j})}.$$

Настроить K набор параметров $w_1 \ldots w_K$ и итоговый алгоритм определим как

$$a(x) = \arg\max_{k\in\{1,..K\}} \langle w_k, x\rangle$$

Рассмотрим след функцию потерь:

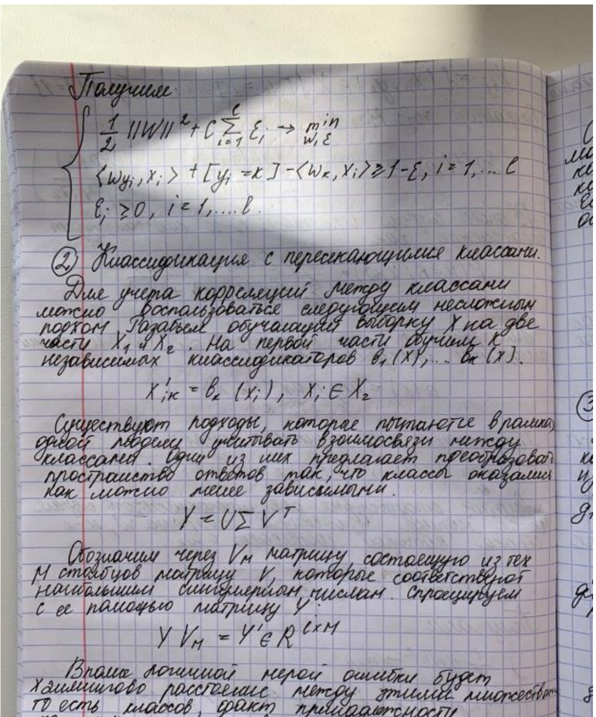
$$\max_k \{\langle w_k, x\rangle + 1 - [k = y(x)]\} - \langle w_{y(x)}, x\rangle$$

Аналог отступствия для многоклассового случая яв-ся норма Фробениуса матрица W, k-я строка которой совпадает с w_k:

$$\rho = \frac{1}{\|W\|^2} = \frac{1}{\sum_{k=1}^{K}\sum_{j=1}^{d} w_{kj}^2}$$

Получаем:

$$\begin{cases} \frac{1}{2}\|W\|^2 \to \min\limits_{W} \\ \langle w_{y_i}, x_i\rangle + [y_i = k] - \langle w_k, x_i\rangle \geq 1, \; i = 1 .. \ell \end{cases}$$



Получим

$$\begin{cases} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{\ell}\varepsilon_i \to \min\limits_{w,\varepsilon} \\ \langle w_{y_i}, x_i\rangle + [y_i = k] - \langle w_k, x_i\rangle \geq 1 - \varepsilon_i, \; i = 1, \ldots \ell \\ \varepsilon_i \geq 0, \; i = 1, \ldots \ell. \end{cases}$$

② Классификация с пересекающимися классами.

Для учета корреляций между классами можно воспользоваться следующим несложным подходом. Разобьем обучающую выборку X на две части X_1 и X_2. На первой части обучим K независимых классификаторов $b_1(x), \ldots b_K(x)$.

$$x'_{ik} = b_k(x_i), \quad x_i \in X_2$$

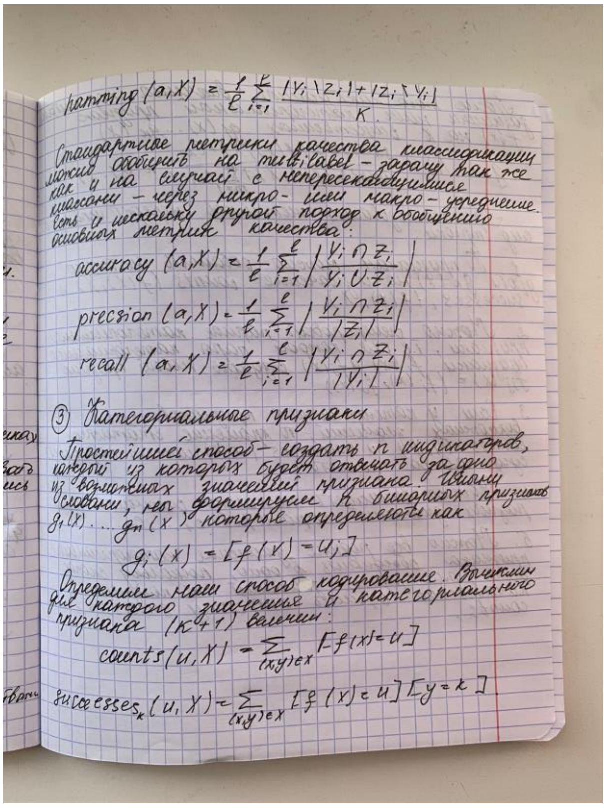
Существуют подходы, которые пытаются в рамках одной модели учитывать взаимосвязи между классами. Один из них предлагает преобразовать пространство ответов так, что классы оказались как можно менее зависимыми.

$$Y = U\Sigma V^T$$

Обозначим через V_M матрицу, состоящую из тех M столбцов матрицы V, которые соответствуют наибольшим сингулярным числам. Спроецируем с ее помощью матрицу Y:

$$YV_M = Y' \in R^{\ell \times M}$$

В роли логичной меры ошибки будет Хэммингово расстояние между этими множествами, то есть классов, факт принадлежности которым угадан неверно;



$$hamming(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{|Y_i \setminus Z_i| + |Z_i \setminus Y_i|}{K}.$$

Стандартные метрики качества классификации можно обобщить на multilabel-задачу так же как и на случай с непересекающимися классами — через микро- или макро- усреднение. Есть и несколько другой подход к обобщению основных метрик качества:

$$accuracy(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left|\frac{Y_i \cap Z_i}{Y_i \cup Z_i}\right|$$

$$precision(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left|\frac{Y_i \cap Z_i}{|Z_i|}\right|$$

$$recall(a,X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left|\frac{Y_i \cap Z_i}{|Y_i|}\right|$$

③ Категориальные признаки

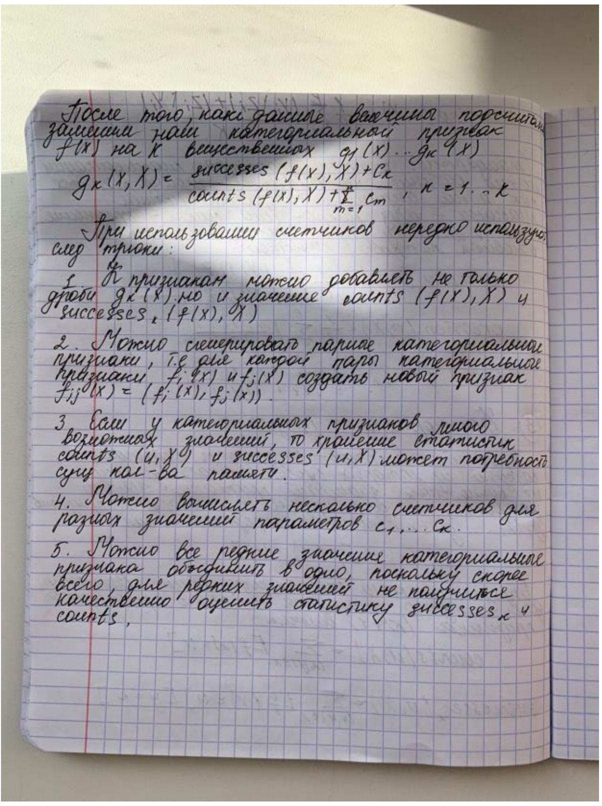
Простейший способ — создать n индикаторов, каждый из которых будет отвечать за одно из возможных значений признака. Иными словами, мы формируем n бинарных признаков $g_1(x) \dots g_n(x)$ которые определяются как

$$g_i(x) = [f(x) = u_i]$$

Определим наш способ кодирования. Вычислим для каждого значения u категориального признака $(K+1)$ величин:

$$counts(u,X) = \sum_{(x,y)\in X}[f(x)=u]$$

$$successes_k(u,X) = \sum_{(x,y)\in X}[f(x)=u][y=k].$$



После того, как данные величины подсчитаны, заменим наш категориальный признак $f(x)$ на K вещественных $g_1(x) \dots g_K(x)$

$$g_k(x,X) = \frac{successes(f(x),X)+c_k}{counts(f(x),X)+\sum_{m=1}^{K}c_m},\ k = 1..K$$

При использовании счетчиков нередко используют след. трюки:

1. К признакам можно добавлять не только дроби $g_k(x)$, но и значение $counts(f(x),X)$ и $successes_k(f(x),X)$

2. Можно сгенерировать парные категориальные признаки, т.е для каждой пары категориальных признаков $f_i(x)$ и $f_j(x)$ создать новый признак $f_{ij}(x) = (f_i(x), f_j(x))$.

3. Если у категориальных признаков много возможных значений, то хранение статистик $counts(u,X)$ и $successes(u,X)$ может потребовать сущ. кол-ва памяти.

4. Можно вычислять несколько счетчиков для разных значений параметров $c_1, \dots c_K$.

5. Можно все редкие значения категориального признака объединить в одно, поскольку скорее всего, для редких значений не получится качественно оценить статистику $successes_k$ и $counts$.