Задача 1. В случае, когда обучающая выборка содержит очень мало объектов, k-fold с небольшим k не подойдет, так как при нем обучение будет вестись на части и без того очень маленькой выборки. В этом случае лучше использовать Leave-one-out-CV, так как при обучении выкидывается всего один объект и проверка проводится на нем. В итоге каждый объект выборки поучаствует в контроле ровно один раз. То есть при обучении мы теряем всего лишь один элемент выборки, а так как выборка маленькая, то обучение N-1 раз не потребует много времени.

В случае же, когда обучающая выборка очень велика, Leave-one-out не подойдет, так как это потребует очень много времени на обучение N-1 раз. В этом случае лучше использовать k-fold. При этом обучение будет производиться k раз (k мало), а так как объектов в выборке очень много, качество обучения не сильно пострадает при обучении на $\frac{k-1}{k}$ объектов.

Задача 2. $\rho_p(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$

При $n=2$, $\rho_p(x,y) = (|x_1-y_1|^p + |x_2-y_2|^p)^{\frac{1}{p}}$, $f_p(x) = \rho_p(x,0) = (|x_1|^p + |x_2|^p)^{\frac{1}{p}}$

Евклидова метрика: $f_2(x) = \rho_2(x,0) = \sqrt{x_1^2 + x_2^2}$

Манхэттенское расстояние: $f_1(x) = |x_1| + |x_2|$

Метрика Чебышева: $f_\infty(x) = \lim_{n\to\infty}(x_1^n + x_2^n)^{\frac{1}{n}}$

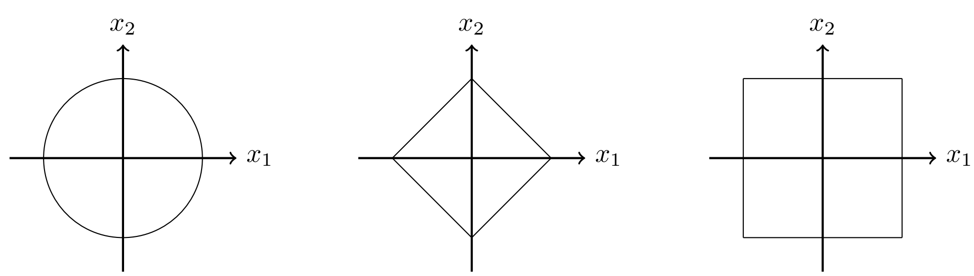
Покажем, что $\lim_{n\to\infty}(x_1^n + x_2^n)^{\frac{1}{n}} = \max\{x_1, x_2\}$

БОО, пусть $\max\{x_1,x_2\} = x_1$, тогда $\lim_{n\to\infty}(x_1^n + x_2^n)^{\frac{1}{n}} \geq \lim_{n\to\infty}(x_1^n)^{\frac{1}{n}} = \lim_{n\to\infty} x_1 = x_1 = \max\{x_1,x_2\}$

С другой стороны, $\lim_{n\to\infty}(x_1^n + x_2^n)^{\frac{1}{n}} \leq \lim_{n\to\infty}(x_1^n + x_1^n)^{\frac{1}{n}} = \lim_{n\to\infty}(2x_1^n)^{\frac{1}{n}} = x_1 \lim_{n\to\infty} 2^{\frac{1}{n}} = x_1 = \max\{x_1,x_2\}$

Таким образом, получатся следующие графики:

Евклидова метрика   Манхэттенское расстояние   Метрика Чебышева



Задача 3.

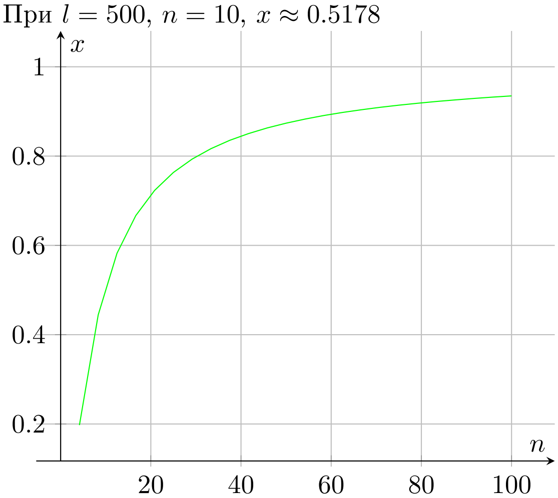
Пусть ξ - минимальное расстояние от центра шара до точки, тогда $F_\xi(x) = P(\xi \leq x) = P($в шаре радиуса x есть хотя бы одна точка$) = 1 - P($все точки вне шара$)$. Вероятность того, что точка находится вне шара радиуса x можно получить если вычесть из объема единичного шара объем шара радиуса x и поделить это на объем единичного шара, то есть $P = \frac{V_1 - V_x}{V_1} = 1 - \frac{V_x}{V_1}$. Так как точки распологаются независимо, вероятность того, что все точки находятся вне шара будет равна $(1 - \frac{V_x}{V_1})^l$.

Объем n-мерного шара радиуса x равен $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} x^n$. Подставим в формулу:

$P($в шаре радиуса x есть хотя бы одна точка$) = 1 - (1 - x^n)^l$

Так как x - медиана, то $F_\xi(x) = 0.5$, $1 - (1-x^n)^l = 0.5$, $(1-x^n)^l = \frac{1}{2}$, $1 - x^n = 2^{-\frac{1}{l}}$, $x = \sqrt[n]{1 - 2^{-\frac{1}{l}}}$.

При $l = 500$, $n = 10$, $x \approx 0.5178$

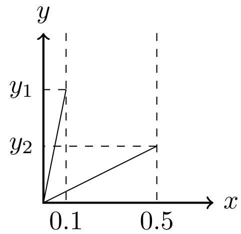


Видно, что при увеличении размерности x быстро растет.

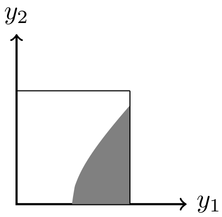
Возьмем теперь значение x=0.5 (будем считать, что мы побороли проклятие размерности при таком значении). Тогда $l = -\frac{1}{\log_2(1-2^{-n})}$.

Возьмем n=100. Для такой размерности чтобы побороть проклятие размерности нужно $9.7 \cdot 10^{13}$ точек, то есть очень много. Таким образом, данная формула наглядно демонстрирует проклятие размерности.

Задача 4.



$y_1, y_2 \in U([0,1])$. Нужно найти вероятность того, что $0.5^2 + y_2^2 < 0.1^2 + y_1^2$, то есть $P(y_1^2 - y_2^2 > 0.24)$



Нужно найти площадь под графиком $y_2 = \sqrt{y_1^2 - 0.24}$, для этого достаточно найти $\int_{\sqrt{0.24}}^{1} \sqrt{x^2 - 0.24}dx \approx 0.275$

Задача 5.

В данном случае число K - гиперпараметр метода KNN. Параметры KNN - координаты объектов из обучающей выборки. В результате у линейной модели на рисунке 1 всего 2 параметра, а у KNN намного больше.

Задача 6.

Вероятность получить на объекте x_i из выборки ответ 1 равна $p(x_i)^{y_i}$, вероятность получить 0 - $(1-p(x_i))^{1-y_i}$. Таким образом, функция правдоподобности примет вид

$L = \prod_{i=1}^{l} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$, прологарифмируем:

$logL = \sum_{i=1}^{l}(y_i \log p(x_i) + (1-y_i)\log(1-p(x_i)))$, что равно формуле для LogLoss со знаком минус.

Задача 7.

1. $(\frac{1}{N}\sum_{i=1}^{N}(y_i - y)^2)' = (\frac{1}{N}(\sum_{i=1}^{N} y_i^2 - 2\sum y_i y + \sum y^2))' = \frac{1}{N}(-2\sum y_i + 2Ny) = -\frac{2}{N}\sum y_i + 2y = 0$, $y = \frac{1}{N}\sum y_i$. То есть нужно взять выборочное среднее. При этом, если мы возьмем величину меньше выборочного среднего, производная будет отрицательной, а если возьмем больше - положительной, следовательно, это точка минимума.

2. $(\frac{1}{N}\sum_{i=1}^{N}|y_i - y|)' = \frac{1}{N}\sum \mathrm{sign}(y - y_i) = 0$. Производная будет равна 0 если взять медиану выборки. При этом если взять величину меньше медианы, производная будет отрицательной (в сумме знаков будет больше значений -1 чем 1), а если больше - положительной, следовательно, это точка минимума.

3. $(-\frac{1}{N}\sum_{i=1}^{N}(y_i \log y + (1-y_i)\log(1-y)))' = -\frac{1}{N}\sum y_i \frac{1}{y} - (1-y_i)\frac{1}{1-y} = 0$

$\sum \frac{y_i(1-y)-(1-y_i)y}{y(1-y)} = 0$, $\sum \frac{y_i - y}{y(1-y)} = \sum(\frac{y_i}{y(1-y)} - \frac{1}{1-y}) = 0$, $\frac{N}{1-y} = \sum \frac{y_i}{y(1-y)}$, $y = \frac{1}{N}\sum y_i$. То есть нужно взять выборочное среднее. При этом производная равна $-\frac{1}{y(1-y)}(\frac{1}{N}\sum y_i - y)$. Так как y_i равны 0 или 1, y будет лежать межну 0 и 1, то есть $-\frac{1}{y(1-y)}$ всегда будет отрицательным. Разность в скобках будет положительной если y меньше выборочного среднего и отрицательной если меньше. То есть в данной точке действительно достигается минимум функции.

Задача 8.

1. Выборочная дисперсия для выборки размера l равна $\frac{1}{l}\sum_{i=1}^{l}(y_i - \overline{y})^2$

При $L(y_i, \overline{y}) = (y_i - \overline{y})^2$, $\Phi(U) = \frac{1}{l}\sum_{i=1}^{l}(y_i - \overline{y})^2$, то есть выборочной дисперсии.

Задача 9.

1. $E\frac{1}{N}\sum_{i=1}^{N}[y_i \neq a(x_i)] = \frac{1}{N}\sum_{i=1}^{N} E[y_i \neq a(x_i)] = \frac{1}{N}\sum_{i=1}^{N}(1 - p_{my_i})$

Разобъем по классам: $= \sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}[y_i = k](1 - p_{mk}) = [\frac{1}{N}\sum_{i=1}^{N}[y_i = k] = p_{mk}] = \sum_{k=1}^{K} p_{mk}(1 - p_{mk})$, что равно индексу Джини.