# Прикладные задачи анализа данных

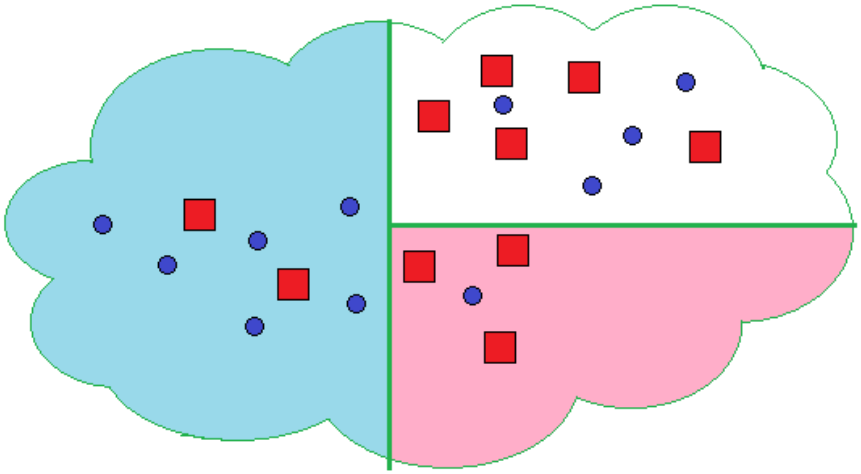
## Минимизация ошибок

**Дьяконов А.Г.**

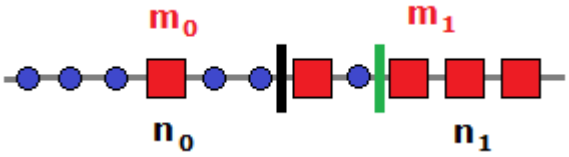
**Московский государственный университет имени М.В. Ломоносова (Москва, Россия)**

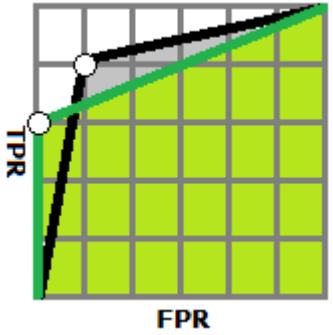
# Использование «тайных знаний» на практике

## Строим деревья в решающем лесе – хотим минимизировать AUC ROC



# Хотим выбирать оптимальный порог




**$n_i$ – числа точек в листах**

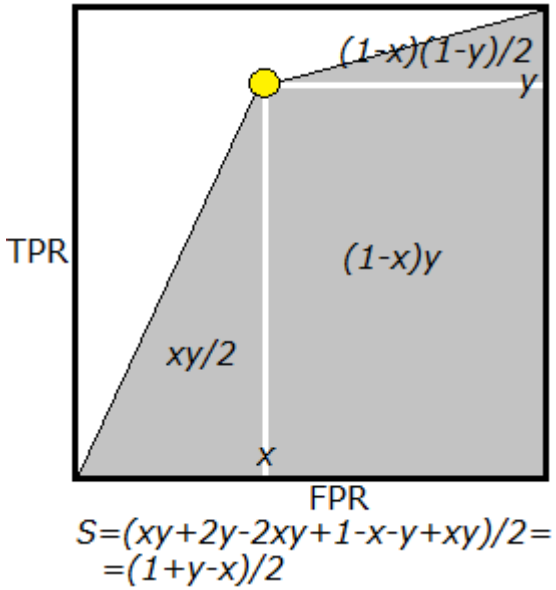
**$m_i$ – числа объектов первого класса в листах**

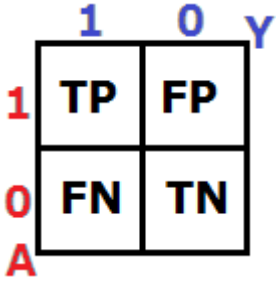
$$m = m_1 + m_0$$

$$n = n_1 + n_0$$

# Уже было…

**Если «вдруг»** $a \in \{0,1\}^q$, $y \in \{0,1\}^q$, $F = AUC$**?! (Сбербанк)**




**Из рисунка**

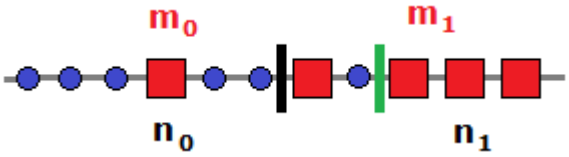
$$AUC = (1 + TPR - FPR)/2 =$$
$$= \frac{1}{2}\left[1 + \frac{TP}{TP+FN} - \frac{FP}{FP+TN}\right]$$

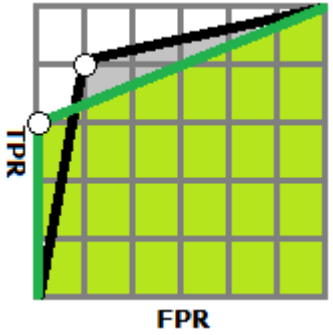
$$= \frac{1}{2}\left[1 + \frac{\|a \cdot y\|}{\|y\|} - \frac{\|a \cdot \bar{y}\|}{\|\bar{y}\|}\right] =$$
$$= \frac{1}{2}\left[\frac{\|a \cdot y\|}{\|y\|} + \frac{\|\bar{a} \cdot \bar{y}\|}{\|\bar{y}\|}\right]$$

**т.е. это точность с оглядкой на мощности классов…**

$$TPR - FPR \to \max$$

# Хотим выбирать оптимальный порог




$$AUC = \frac{1}{2}\left[\frac{m_1}{m} + \frac{n_0 - m_0}{n-m}\right] =$$

$$= \frac{1}{2}\left[\frac{m_1}{m} + \frac{(n-m)-(n_1-m_1)}{n-m}\right] =$$

$$= \frac{1}{2} + \frac{1}{2}\left[\frac{m_1}{m} - \frac{n_1-m_1}{n-m}\right]$$

# Хотим выбирать оптимальный порог

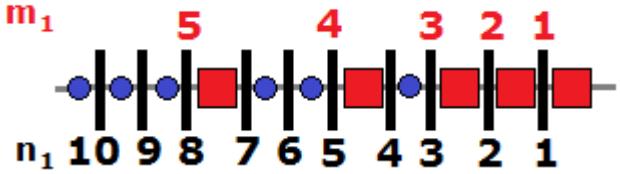
**Логично** $|AUC - 0.5| \to \max$

$$\left|\frac{m_1}{m} - \frac{n_1 - m_1}{n - m}\right| \to \max$$

$$\left|\frac{m_1 n - n_1 m}{m(n-m)}\right| \to \max$$

$$|m_1 n - n_1 m| \to \max$$

**А ведь тогда просто реализовать перебор порогов в скриптовых языках**

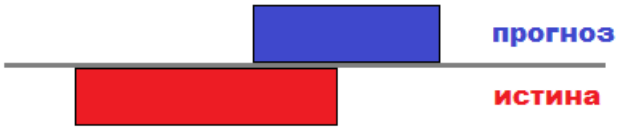


# RF для AUC

**Получили «новую» модель алгоритмов!**

**Это из старых записей…**

# Задачи с интервальными признаками…
# Как решать




**Качество измеряем, например так:**

$$\frac{|A \cap B|}{|A \cup B|}$$

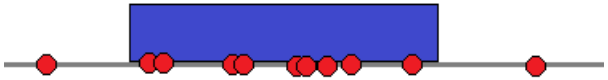
# 1 способ

**Две задачи:**
**Целевой признак – начало интервала,**
**Целевой признак – конец интервала**

**- на практике работает не очень хорошо**
**- надо дорабатывать классические алгоритмы**
**(т.к. в случае начала интервала лучше занижать…)**

# 2 способ

**Целевой признак – середина интервала,**
**плюс оцениваем отклонение от середины**



**- иногда противоречит природе данных**
**(интервал заходит в отрицательную область)**

# Концепция решающего правила

## Как всё-таки минимизировать нужный функционал…

1. **Есть предварительный ответ**

$$[a,b]$$

2. **Формируем окончательный параметрический…**

$$\left[\frac{a+b}{2}-\varepsilon\frac{b-a}{2},\frac{a+b}{2}-\varepsilon\frac{b+a}{2}\right]$$

3. **Настраиваем параметр**

**Прямой перебор – явная минимизация**

# Можно и по-другому…

**Но:**

1. **Есть базовые алгоритмы (операторы)**
2. **Есть параметризованный способ перевода их ответов в нужные**
3. **Прямая минимизация функционала**

# Из задачи Rossmann Store Sales

## Root Mean Square Percentage Error (RMSPE)

$$\sqrt{\frac{1}{|\{i \mid y_i > 0\}|}\sum_{i:\, y_i>0}\left(\frac{a_i - y_i}{y_i}\right)^2}$$ и

## Оправдание деформации логарифмом…

# Оправдание деформации логарифмом...

**Ищем деформацию**

$$\frac{a-y}{y} \approx F(a)-F(y)$$

**чтобы функционал превратился в RMSE**

$$\sqrt{\frac{1}{|\{i \mid y_i > 0\}|}\sum_{i:\, y_i>0}(F(a_i)-F(y_i))^2}$$

**Пусть** $a = y+\delta$**, тогда**

$$\frac{\delta}{y} \approx F(y+\delta)-F(y) = F'\delta + o(\delta)$$

**решим уравнение**

$$\frac{\delta}{y} = F'\delta$$

# Оправдание деформации логарифмом…

$$\frac{1}{y} = \frac{\partial F}{\partial y}$$

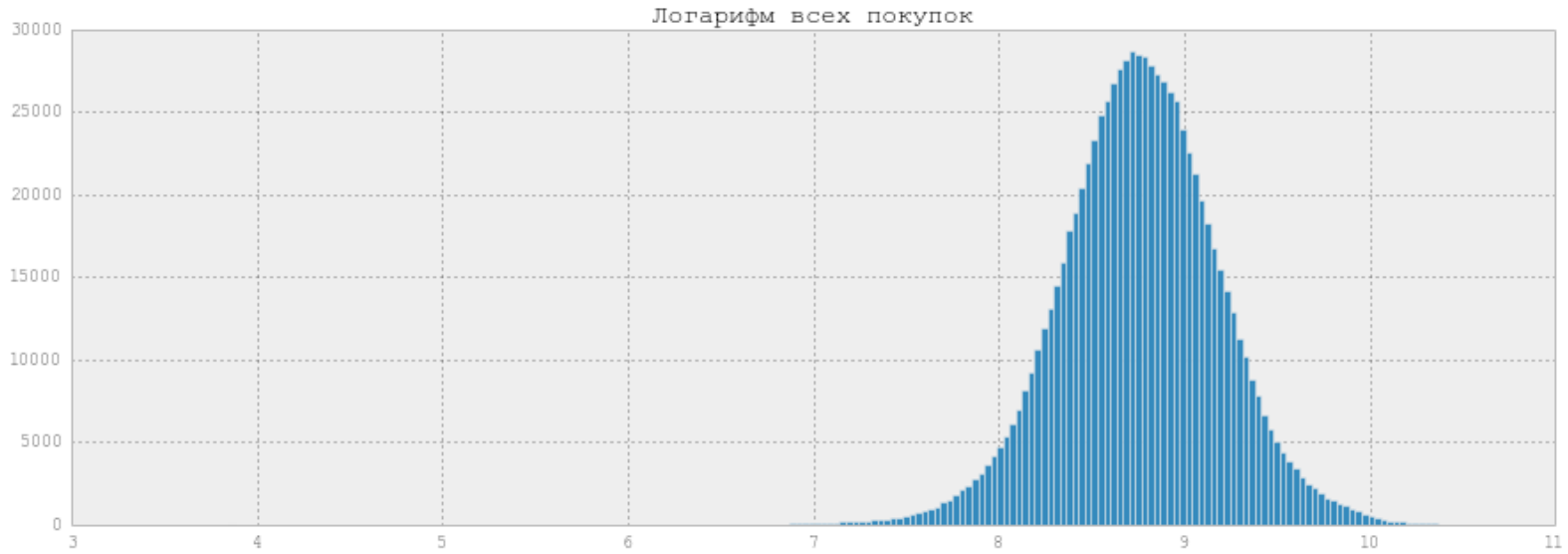
$$F(y) = \ln|y| + C$$

**Выбираем деформацию** $F(y) = \ln|y|$

**Но, возможно, всё проще…**

**при логарифмировании отклонения похожи на нормальные**

# Распределения покупок

# Метод градиентного спуска

## Задача оптимизации

$$J(\tilde{w}) \to \min$$

**пусть это**

**ФУНКЦИЯ ОШИБКИ (ПАРАМЕТРЫ АЛГОРИТМА)**

$$\tilde{w} := \tilde{w} - \alpha \left. \frac{\partial J}{\partial \tilde{w}} \right|_{\tilde{w}}$$

# Возьмём конкретную задачу и метод

**Качество: LOG LOSS**

**Метод: логистическая регрессия**

**(правильнее: сигмоида!)**

$$LOGLOSS = -\frac{1}{q}\sum_{i=1}^{q}(y_i \log a_i + (1-y_i)\log(1-a_i))$$

$$a = \frac{1}{1+e^{-z}}$$

$z = z(w)$ **– может как-то зависеть от параметров** $w$.

**На конкретном объекте:**

$$J(w) = -\begin{cases} \log a, & y = 1, \\ \log(1-a), & y = 0. \end{cases}$$

**Итак,**

$$J(w) = -\begin{cases} \log\left(\dfrac{1}{1+e^{-z}}\right), & y = 1, \\ \log\left(1 - \dfrac{1}{1+e^{-z}}\right), & y = 0. \end{cases}$$

$$J(w) = -\begin{cases} -\log\left(1+e^{-z}\right), & y = 1, \\ -z - \log\left(1+e^{-z}\right), & y = 0. \end{cases}$$

$$\frac{\partial \log\left(1+e^{-z}\right)}{\partial w} = -\frac{1}{1+e^{-z}} e^{-z} \frac{\partial z}{\partial w}$$

$$\frac{\partial J(w)}{\partial w} = -\frac{\partial z}{\partial w}\begin{cases} \dfrac{e^{-z}}{1+e^{-z}}, & y = 1, \\ -1 + \dfrac{e^{-z}}{1+e^{-z}}, & y = 0. \end{cases}$$

**Поэтому**

$$\frac{\partial J(w)}{\partial w} = -\frac{\partial z}{\partial w}\begin{cases} 1-\dfrac{1}{1+e^{-z}}, & y=1, \\ 0-\dfrac{1}{1+e^{-z}}, & y=0. \end{cases} = -\frac{\partial z}{\partial w}(y-a)$$

**Получаем формулу для коррекции весов:**

$$w = w + \alpha(y-a)\frac{\partial z}{\partial w}$$

**Очень логичная: изменение зависит от величины ошибки**

$$(y-a)$$

**В классической логистической регрессии**

$$a = \frac{1}{1 + e^{-\sum_{t=1}^{n} w_t [x]_t}}$$

**(линейная комбинация признаков)**

**Поэтому**

$$w = w + \alpha(y - a)x$$

$x$ **– признаковое описание объекта**

# Вопрос с подвохом

**Качество: LOG LOSS**

**Метод: линейная регрессия**

$$J(w) = -\begin{cases} \log(z), & y = 1, \\ \log(1-z), & y = 0. \end{cases}$$

$$\frac{\partial J(w)}{\partial w} = -\frac{\partial z}{\partial w}\begin{cases} 1/z, & y = 1, \\ -1/(1-z), & y = 0, \end{cases} = \frac{1}{z+y-1}\frac{\partial z}{\partial w}$$

**тогда**

$$w = w - \frac{\alpha}{z+y-1}\frac{\partial z}{\partial w}$$

**Что смущает в этой формуле?**

**Почему так получилось?**

# Вопрос с подвохом

$$w = w - \frac{1}{z + y - 1}\frac{\partial z}{\partial w}$$

**Коррекция происходит даже при абсолютно правильном ответе…**

минимум не при z=1

$$J(w) = -\begin{cases} \log(z), & y = 1, \\ \log(1-z), & y = 0. \end{cases}$$

**Нужны ещё ограничения**

$$\min(\max(z, 0), 1)$$

**В логистической регрессии**

$$\frac{1}{1+e^{-z}} \in [0, 1]$$

# Линейная регрессия с НСКО

$$J(\tilde{w}) = (\tilde{w}^{\mathrm{T}} \cdot \tilde{x} - y)^2 \to \min$$

$\tilde{x}$ **– объект,**

$y$ **– его регрессионная метка**

$$\frac{\partial J}{\partial \tilde{w}} = 2 \cdot (\tilde{w}^{\mathrm{T}} \cdot \tilde{x} - y) \cdot \tilde{x}$$

$$\tilde{w} := \tilde{w} - \alpha \cdot (\tilde{w}^{\mathrm{T}} \cdot \tilde{x} - y) \cdot \tilde{x}$$

**Выберем** $\alpha$

**Коррекция такая же как в логистической регрессии с logloss-ом!**

# Метод наискорейшего спуска

$$((\tilde{w}-\alpha\cdot(\tilde{w}^{\mathrm{T}}\cdot\tilde{x}-y)\cdot\tilde{x})^{\mathrm{T}}\cdot\tilde{x}-y)^2\to\min$$

$$\tilde{w}^{\mathrm{T}}\cdot\tilde{x}-\alpha\cdot(\tilde{w}^{\mathrm{T}}\cdot\tilde{x}-y)\cdot\tilde{x}^{\mathrm{T}}\cdot\tilde{x}-y=0$$

$$\tilde{w}^{\mathrm{T}}\cdot\tilde{x}-y=\alpha\cdot(\tilde{w}^{\mathrm{T}}\cdot\tilde{x}-y)\cdot\tilde{x}^{\mathrm{T}}\cdot\tilde{x}$$

$$\alpha=\frac{1}{\tilde{x}^{\mathrm{T}}\cdot\tilde{x}}$$

# Задача 1.

**Качество:** СКО

$$J = \frac{1}{q}\sum_{i=1}^{q}(y_i - a_i)^2$$

**Метод:** логистическая регрессия

$$a = \frac{1}{1+e^{-z}}$$

**Вычислить формулу для коррекции весов методом стохастического градиентного спуска**

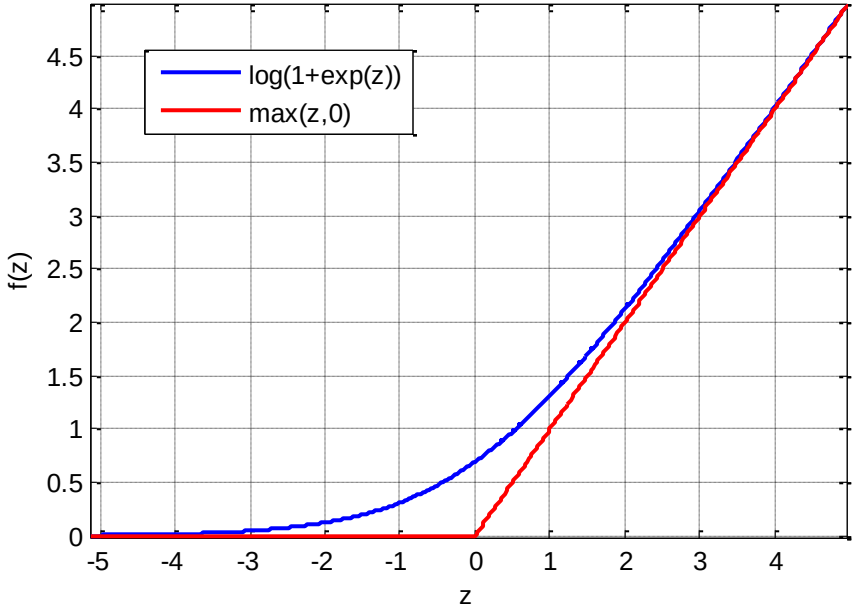
# Задача 2.

**Качество: СКО**

$$J = \frac{1}{q}\sum_{i=1}^{q}(y_i - a_i)^2$$

**Метод:** $a = \ln(1+e^z)$

**Вычислить формулу для коррекции весов методом стохастического градиентного спуска**



# Задача 2. Ответ

$$w = w - \alpha \frac{(a-y)}{1+e^{-z}} \frac{\partial z}{\partial w}$$

**Почти классический вариант (линейная регрессия + СКО), но с поправкой на отрицательной оси…**

$$w = w - \alpha \frac{(a-y)}{1+e^{-z}} \frac{\partial z}{\partial w}$$

классика

сигмоида

$$\frac{(a-y)}{1+e^{-z}} \approx \begin{cases} (a-y), & z >> 0 \\ 0, & z << 0 \end{cases}$$

**Всё очень логично!**

# Задача 1. Ответ

$$w = w - \alpha \cdot a(a-1)(a-y)\frac{\partial z}{\partial w}$$

что-то новое...

$$w = w - \alpha \cdot a(a-1)(a-y)\frac{\partial z}{\partial w}$$

классика

**Вопрос:** что плохого в это формуле?

# Задача 1. Ответ

$$w = w - \alpha \cdot a(a-1)(a-y)\frac{\partial z}{\partial w}$$

что-то новое...

$$w = w - \alpha \cdot a(a-1)(a-y)\frac{\partial z}{\partial w}$$

классика

**Вопрос: что плохого в это формуле?**

**В случае полностью неправильного ответа, например**

$$y = 0, a \approx 1$$

**коррекции почти не будет:**

$$a(a-1)(a-y) \approx 0$$

**Вопрос: что с этим делать?**