Машинное обучение в экономике

Метод ближайших соседей

Потанин Богдан Станиславович

доцент, кандидат экономических наук

2024–2025

- Методы классификации:
  - Метод ближайших соседей.
- Базовые понятия:
  - Метрики и индексы расстояний.
  - Нормализация признаков.
  - Виды прогнозов.
  - Матрица ошибок (путаницы).
  - Подбор оптимального порога вероятностей для прогнозирования.
  - Точность, полнота, F1-метрика, средняя точность по классам.
  - ROC-кривая и AUC.
  - Выигрыш (gain) и подъем (lift).
  - Гиперпараметры.
  - Несбалансированная выборка и балансировка классов.

Метод ближайших соседей

Интуиция

- Чтобы спрогнозировать, совершит ли покупку тот или иной индивид, мы можем изучить действия, совершавшиеся похожими на него покупателями.
- **Вопрос** – как определить схожих покупателей?
- **Ответ** – на основании меры сходства признаков этих покупателей: возраст, образование и т.д.
- **Идея** – если покупатели со схожими признаками совершали покупку, то мы предполагаем, что и данный покупатель также совершит покупку.
- **Проблема** – необходимо каким-то образом измерить сходство между покупателями.
- **Решение** – воспользоваться метриками расстояния, позволяющими измерить сходство между покупателями в зависимости от их характеристик (признаков): доход, возраст и т.д.

- Для измерения расстояния между двумя m-мерными векторами x и y используется метрика, представляющая собой функцию $d(x, y)$, удовлетворяющую следующим свойствам:
  - **Неотрицательность**: $d(x, y) \geq 0$.
  - **Тождественность**: $d(x, y) = 0$ тогда и только тогда, когда $x = y$.
  - **Симметричность**: $d(x, y) = d(y, x)$.
  - **Неравенство треугольника**: $d(x, y) \leq d(x, z) + d(y, z)$ для любого z.
- Наиболее популярные метрики расстояния:
  - **Минковского**: $d(x, y) = \left(\sum_{i=1}^{m} (|x_i - y_i|)^{\lambda} \right)^{\frac{1}{\lambda}}$, где $\lambda \in N$

    Чем больше λ, тем сильнее штрафуются существенные различия между x_i и y_i.
  - **Евклидова**: $(\lambda = 2)$: $d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$

    Пример: $x = (4, 2, 8)$, $y = (0, 5, 8)$, $d(x, y) = \sqrt{(4-0)^2 + (2-5)^2 + (8-8)^2} = 5$
  - **Манхэттен**: $(\lambda = 1)$: $d(x, y) = \sum_{i=1}^{m} |x_i - y_i|$

    Пример: $x = (4, 2, 8)$, $y = (0, 5, 8)$, $d(x, y) = |4-0| + |2-5| + |8-8| = 7$

- **Идея** – прогнозируем наблюдению значение целевой переменной, совпадающее со значением целевой переменной другого, наиболее близкого к нему, с точки зрения признаков, наблюдения, именуемого **ближайшим соседом**.
- Через X_i обозначим i-е наблюдение и запишем классифицирующее правило:

$$\hat{y}(x) = Y_i, \text{ где } i = \underset{j \in \{1,\dots,n\}}{\operatorname{argmin}} \; d(x, X_j)$$

- Прогноз $\hat{y}(x)$ это Y_i ближайшего соседа x с точки зрения функции растояния $d()$.
- Рассмотрим выборку $X_1 = (0,1)$, $X_2 = (1,2)$, $X_3 = (0,6)$, $X_4 = (2,2)$, $Y = (1,0,0,1)$ и найдем ближайшего соседа для $x = (1,3)$ с помощью расстояния Евклида:

$$d(x, X_1) = \sqrt{(1-0)^2 + (3-1)^2} = \sqrt{5} \qquad d(x, X_2) = \sqrt{(1-1)^2 + (3-2)^2} = 1$$
$$d(x, X_3) = \sqrt{(1-0)^2 + (3-6)^2} = \sqrt{10} \qquad d(x, X_4) = \sqrt{(1-2)^2 + (3-2)^2} = \sqrt{2}$$

- Ближайшим соседом x является X_2, а значит $\hat{y}(x) = Y_2 = 0$.

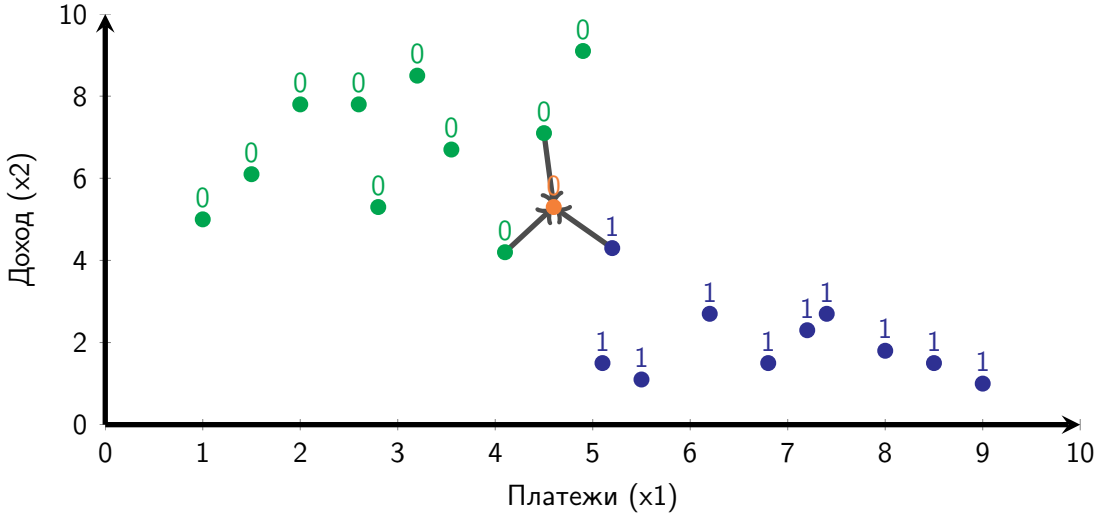
- **Проблема** – прогноз, основанный всего на одном наблюдении, будет обладать очень большой дисперсией, что мотивирует учет большего числа соседей.
- **Решение** – по аналогии с одномерным случаем находим k ближайших соседей и выбираем наиболее часто встречающееся среди них значение целевой переменной.
- Число соседей k часто выбирается нечетным, для того, чтобы избежать ситуации, когда 0 и 1 поровну среди соседей.
- Рассмотрим выборку $X_1 = (0, 1)$, $X_2 = (1, 2)$, $X_3 = (0, 6)$, $X_4 = (2, 2)$, $Y = (1, 0, 0, 1)$ и найдем $k = 3$ ближайших соседа для $x = (1, 3)$ с помощью дистанции Манхэттен:

$$d(x, X_1) = |1 - 0| + |3 - 1| = 3 \qquad d(x, X_2) = |1 - 1| + |3 - 2| = 1$$
$$d(x, X_3) = |1 - 0| + |3 - 6| = 4 \qquad d(x, X_4) = |1 - 2| + |3 - 2| = 2$$

- Ближайшими соседями x являются X_1, X_2 и X_4.
- Поскольку $Y_1 = 1$, $Y_2 = 0$ и $Y_4 = 1$, то $\hat{y}(x) = 1$.

Метод ближайших соседей

Графическая иллюстрация идеи на примере дефолта



- При поиске ближайших соседей иногда возникает ситуация, при которой расстояние сразу до нескольких соседей является минимальным. Будем именовать таких ближайших соседей **спорными**.
- Для простоты удобно помыслить пример, в котором $k = 1$ и X_1, X_2 оба являются ближайшими соседями x, то есть, в частности, $d(x, X_1) = d(x, X_2)$.
- В таком случае возможны различные варианты, в том числе:
  - Выбрать первого из спорных ближайших соседей, то есть с наименьшим индексом. В рассматриваемом примере будет выбран сосед X_1, а не X_2, в связи с тем, что $1 < 2$. **Важно** – если не сказано иное, в рамках курса при решении задач необходимо применять этот способ.
  - Выбрать случайным образом (например, равновероятно) одного из спорных ближайших соседей.
  - Увеличивать k до тех пор, пока не останется спорных ближайших соседей. В рассматриваемом примере достаточно положить $k = 2$.

- Обозначим через $N_{y,1}, ..., N_{y,k}$ значения целевых переменных ближайших соседей x.
- Условная вероятность может быть оценена как доля ближайших соседей с соответствующим значением целевой переменной:

$$\hat{P}(Y_i = 1 | X_i = x) = \frac{1}{k} \sum_{i=1}^{k} N_{y,i} = \frac{N_Y}{k}$$

- В предыдущем примере $k = 3$ и целевые переменные ближайших соседей равнялись $Y_1 = 1,\ Y_2 = 0,\ Y_4 = 1$, откуда $N_Y = 2$, а значит:

$$\hat{p}_x = \hat{P}(Y_i = 1 | X_i = x) = \frac{2}{3}$$

- Как и ранее прогнозируемое значение может зависеть от порога:

$$\hat{y}(x) = I(\hat{p}_x \geq c)$$

Метод ближайших соседей

Взвешенные ближайшие соседи

- **Проблема** – при большом числе соседей k оценки метода ближайших соседей страдают от большого смещения, а при малом количестве соседей – от большой дисперсии.
- **Решение** – взять достаточно много ближайших соседей k, но при построении прогноза учитывать их значения целевой переменной с разными весами, пропорциональными расстояниям до этих соседей.
- Обозначим через $N_{(x,1)}, .., N_{(x,k)}$ признаки ближайших соседей наблюдения с признаками x, а через $N_{(y,1)}, .., N_{(y,k)}$ – значения их целевых переменных.
- **Интуиция** – чем больше расстояние $d(x, N_{(x,i)})$, тем меньший вес должен получить i-й сосед. Этой идее удовлетворяет, например, функция $1/d(x, N_{(x,i)})$.
- Классификатор присваивает значение 1, если суммарный вес соседей с $N_{(y,i)} = 1$ больше, чем суммарный вес соседей с $N_{(y,i)}(y) = 0$:

$$\hat{y}(x) = I\left(\sum_{i:N_{(y,i)}=1} \frac{1}{d(N_{(x,i)}, x)} \geq \sum_{i:N_{(y,i)}=0} \frac{1}{d(N_{(x,i)}, x)}\right)$$

- В общем случае можно ориентироваться на оценки условных вероятностей:

$$\hat{P}\left(Y_i = 1 | X_i = x\right) = \left(\sum_{i:N_{(y,i)}=1} 1/d(N_{(x,i)}, x)\right) / \left(\sum_{i=1}^{k} 1/d(N_{(x,i)}, x)\right)$$

Метод ближайших соседей

Нормализация признаков

- Метод ближайших соседей чувствителен к шкале измерения признаков.
- Например, пусть признаки X_{*1} и X_{*2} отражают возраст индивида и его рост.
- Пусть при росте, измеренном в метрах, $x = (25, 1.8)$, $X_1 = (30, 1.9)$ и $X_2 = (26, 1.6)$.
- Тогда в соответствии с расстоянием Евклида $d(x, X_1) \approx 5$ и $d(x, X_2) \approx 1$, а значит ближайшим соседом x является X_2.
- Если рост измерен в сантиметрах, то $x = (25, 180)$, $X_1 = (30, 190)$ и $X_2 = (26, 160)$, откуда $d(x, X_1) \approx 11.2$ и $d(x, X_2) \approx 20$, а значит ближайшим соседом x является X_1.
- **Проблема** – даже несущественные (существенные) признаки с мелкими (крупными) единицами измерения могут вносить (не)существенный вклад в расчет расстояний.
- **Решение** – осуществить **нормализацию** признаков, то есть привести их к сопоставимой шкале, например, вычтя выборочное среднее и поделив на выборочное стандартное отклонение.
- В качестве альтернативы можно провести нормализацию к шкале $[0, 1]$ вычтя у каждого признака минимальное значение и поделив на разницу между максимальным и минимальным значениями.

Метод ближайших соседей

Индексы расстояний между категориальными признаками

- Интуиция подсказывает, что рассмотренные метрики расстояний хорошо подходят для признаков, измеренных в непрерывной шкале, но могут быть не так хороши для категориальных, в частности, бинарных признаков.
- Если все признаки являются бинарными, то в качестве альтернативы метрике часто используют различные индексы, например, индекс **Рассела-Рао**, рассчитываемый как доля случаев, когда признаки X_i и X_j одновременно равняются 1.
- Например, если $X_i = (1, 0, 1, 1, 0)$ и $X_j = (1, 1, 0, 1, 0)$, то соответствующий индекс будет равен $2/5$.
- Индекс **Сокаля-Минхера** рассчитывается как доля совпадающих значений. В предыдущем примере он окажется равен $3/5$.
- На практике часто встречаются **разреженные данные** (sparse data), в которых большинство значений являются нулями.
- Например, мы можем пытаться спрогнозировать понравится ли новая одежда бренда индивиду, используя информацию о том, какую одежду он покупал ранее. Как правило индивиды покупают лишь небольшое количество одежды из всего доступного ассортимента, из-за чего бинарные переменные на факт покупки той или иной одежды, как правило, будут равняться нулю.
- В таких случаях удобно применять индекс **Жаккара**, рассчитываемый как отношение числа совпадающих 1 к общечему числу признаков за вычетом числа совпадающих нулей, так как в разряженных данных нули совпадают очень часто.

- При большом числе наблюдений n поиск ближайших соседей является весьма сложной с вычислительной точки зрения задачей, требующий расчета расстояния до каждого из наблюдений.
- В качестве альтернативы используются методы, которые находят ближайших соседей не в точности, а приблизительно.
- Одними из наиболее известных алгоритмов, позволяющих приблизительно найти ближайших соседей, именуется **k-d дерево** (k-d tree).
- Обсуждение данных алгоритмов выходит за рамки курса. Общая рекомендация заключается в том, чтобы применять эти алгоритмы в случаях, когда без их использования метод работает слишком медленно, что часто случается при большом числе наблюдений.
- Также, при малом числе ближайших соседей использование приблизительного поиска вместо точного иногда помогает избежать проблемы переобучения.

Оценивание качества прогноза

Виды ошибок прогноза

- Результаты прогнозов бывают следующих видов:
  1. **Верный положительный** (TP – true positive) – когда $y = 1$ предсказан как $\hat{y} = 1$. Пример – модель предсказала дефолт и произошел дефолт.
  2. **Верный отрицательный** (TN – true negative) – когда $y = 0$ предсказан как $\hat{y} = 0$. Пример – модель предсказала отсутствие дефолта, и он не произошел.
  3. **Ложный положительный** (FP – false positive) – когда $y = 0$ предсказан как $\hat{y} = 1$. Пример – модель предсказала дефолт, но он не произошел.
  4. **Ложный отрицательный** (FN – false negative) – когда $y = 1$ предсказан как $\hat{y} = 0$. Пример – модель предсказала отсутствие дефолта, но он произошел.
- Каждый из этих результатов может иметь различную цену (последствия).
- Например, верный отрицательный TN прогноз дефолта может принести умеренную прибыль для банка, поскольку подразумевает, что кредит будет выдан заемщику, который сможет его вернуть.
- В то же время ложный отрицательный FN прогноз дефолта может создать серьезные проблемы, поскольку предполагает выдачу кредита заемщику, не способному его вернуть.
- Поэтому, в классификаторе $I(\hat{p}_x \geq c)$ подбор порогового значения c часто осуществляется исходя из цен (последствий) результатов прогнозов.

Оценивание качества прогноза

Матрица ошибок (confusion matrix)

- Рассмотрим различные результаты прогнозов в форме таблицы:

Прогноз	Истина	Тип прогноза
1	1	TP
0	0	TN
1	0	FP
1	1	TP
0	1	FN
0	1	FN
1	1	TP
0	0	TN

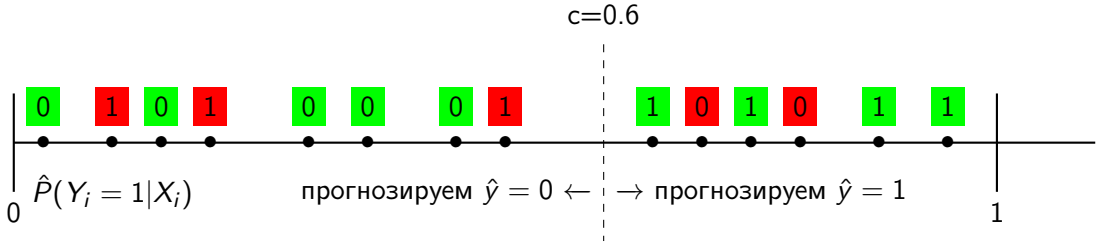
- Число прогнозов различного вида можно представить в форме **матрицы ошибок (путаницы)** (confusion matrix).

	Прогноз = 1	Прогноз = 0
Истина = 1	TP = 3	FN = 2
Истина = 0	FP = 1	TN = 2

- Отметим, что $\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = \frac{3+2}{3+2+1+2} = \frac{5}{8}$.

Оценивание качества прогноза

Графическая иллюстрация связи между порогом и числом прогнозов различного вида



- TP (верный положительный) – зеленые единицы
- TN (верный отрицательный) – зеленые нули
- FP (ложный положительный) – красные нули
- FN (ложный отрицательный) – красные единицы

$$\hat{y}(x) = I(\hat{p}_x \geq c), \text{ где } \hat{p}_x = \hat{P}\left(Y_i = 1|X_i = x\right)$$

Оценивание качества прогноза

Связь между порогом и числом прогнозов различного вида

- Обозначим через TP, TN, FP и FN количество соответствующих прогнозов.
- При увеличении порога c в классификаторе $I(\hat{p}_x \geq c)$ прогнозы $\hat{y} = 1$ возникают реже, что будет приводить к росту TN (хорошо) и FN (плохо), а также к уменьшению TP (плохо) и FP (хорошо).
- В задаче с прогнозированием дефолта не так страшно не выдать кредит хорошему заемщику FP, как выдать кредит плохому заемщику FN.
- Следовательно, в этой задаче мы можем пожертвовать TN прогнозами ради снижения числа FN прогнозов, что мотивирует рассматривать $c < 0.5$.
- Например, при $c = 0.2$ мы прогнозируем дефолт заемщикам, у которых вероятность дефолта равняется хотя бы 0.2, для того, чтобы перестраховаться от крайне нежелательных случаев, когда кредит выдается неплатежеспособному заемщику – снижаем FN (хорошо). При этом мы теряем часть клиентов, которые на самом деле могли бы вернуть кредит – уменьшаем TN (плохо).

- Рассмотрим задачу подбора порога c (threshold) в классификаторе $I(\hat{p}_x \geq c)$.
- Обозначим через p_{TP}, p_{TN}, p_{FP}, p_{FN} цены соответствующих прогнозов. Обычно эти цены отрицательны для ложных прогнозов и положительные для верных.
- **Идея** – перебрать все возможные пороговые значения c и выбрать то, что принесет наибольшую **прибыль** (выгоду) с учетом различных цен результатов прогнозов.
- **Проблема** – поскольку порогов $c \in (0, 1)$ бесконечно много, то какие значения перебирать?
- **Решение проблемы** – отсортировать все спрогнозированные вероятности $\hat{p}_i$ в порядке возрастания и поочередно использовать каждую из них в качестве порога, обозначаемого c_i.
- Через $\mathrm{TP}(c_i)$, $\mathrm{FP}(c_i)$, $\mathrm{TN}(c_i)$, $\mathrm{FN}(c_i)$ обозначим количество соответствующих прогнозов (например, в тестовой выборке), при использовании порога c_i.
- Оптимальное значение порога определяется, например, из решения:

$$c = \underset{c_i}{\operatorname{argmax}}\ p_{\mathrm{TP}}\mathrm{TP}(c_i) + p_{\mathrm{TN}}\mathrm{TN}(c_i) + p_{\mathrm{FP}}\mathrm{FP}(c_i) + p_{\mathrm{FN}}\mathrm{FN}(c_i)$$

Оценивание качества прогноза

Точность и полнота

- Иногда исследователю недоступны цены прогнозов P_{TP}, P_{TN}, P_{FT} и P_{FN}, но при этом интуиция подсказывает, что качество прогнозов разного типа не является равноценным.
- В таком случае используются альтернативные метрики качества модели, например, **точность** (precision) и **полнота** (recall):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{доля верно предсказанных 1 среди всех } \textbf{предсказанных } 1$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{доля верно предсказанных 1 среди всех } \textbf{истинных } 1$$

- Например, в скоринговых моделях доля верно предсказанных дефолтов среди тех, кому мы предсказали дефолт, отражает точность, а среди тех, у кого на самом деле произошел дефолт – полноту.
- Чем выше точность и полнота, тем лучше модель справляется с прогнозами 1. Эта идея используется в F1-**метрике**:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- F1-метрика предполагает, что точность прогноза 1 более важна, чем точность прогноза 0.

- Часто доли 1 и 0 в выборке могут существенно различаться. В таком случае говорят, что данные являются **несбалансированными**.
- Например, клиентов с дефолтом обычно гораздо меньше, чем без дефолта.
- Обозначим через recall(1) и recall(0) полноту, посчитанную для 1 и 0 соответственно. В последнем случае формула будет иметь вид $\text{recall}(0) = \frac{\text{TN}}{\text{TN}+\text{FP}}$.
- **Средняя полнота по классам** (иногда именуется точностью) рассчитывается как:

$$\text{ACA} = \frac{\text{recall}(1) + \text{recall}(0)}{2}$$

- Достигнуть большой средней полноты по классам можно и в случае, когда лишь однин из recall(1) или recall(0) достаточно велик, а другой – гораздо меньше. Для того, чтобы повысить штраф за крайне малое значение одного из этих показателей среднюю полноту по классам часто считают с помощью гармонического среднего:

$$\text{ACA} = \frac{2\left(\text{recall}(1) \times \text{recall}(0)\right)}{\text{recall}(1) + \text{recall}(0)}$$

Оценивание качества прогноза

ROC-кривая

- При подборе оптимального порога в классификаторе $I(\hat{p} \geq c)$ мы можем стремиться найди баланс между следующими долями:

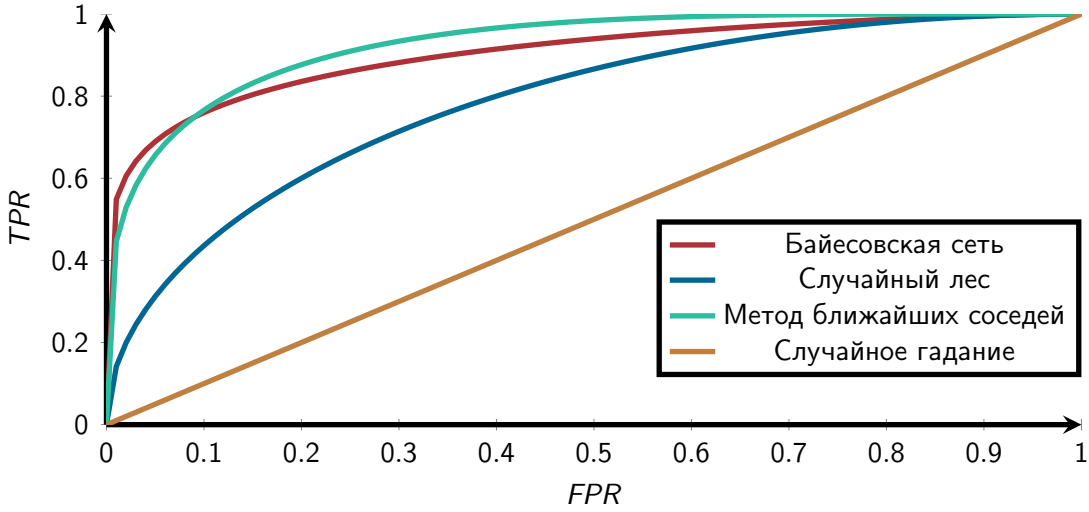
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{доля предсказанных 1 среди всех истинных 1}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \qquad \text{доля предсказанных 1 среди всех истинных 0}$$

- **Важно** – по мере уменьшения порога c растут FPR (false positive rate - плохо) и TPR (true positive rate - хорошо).
- Перебирая различные значения порога (отсортированные оценки условных вероятностей) мы получаем все возможные комбинации TPR и FPR в наших данных.
- **ROC-кривая** отражает график зависимости между TPR и FPR.
- ROC-кривую часто строят сразу для нескольких моделей. В таком случае, если график одной модели в какой-то точке находится над графиком другой модели, то значит, при прочем равном FPR он дает более высокое значение TPR, следовательно эта модель предпочтительна при соответствующем пороге.
- **Применение** – можно понять, при каких порогах различные модели имеют преимущество.

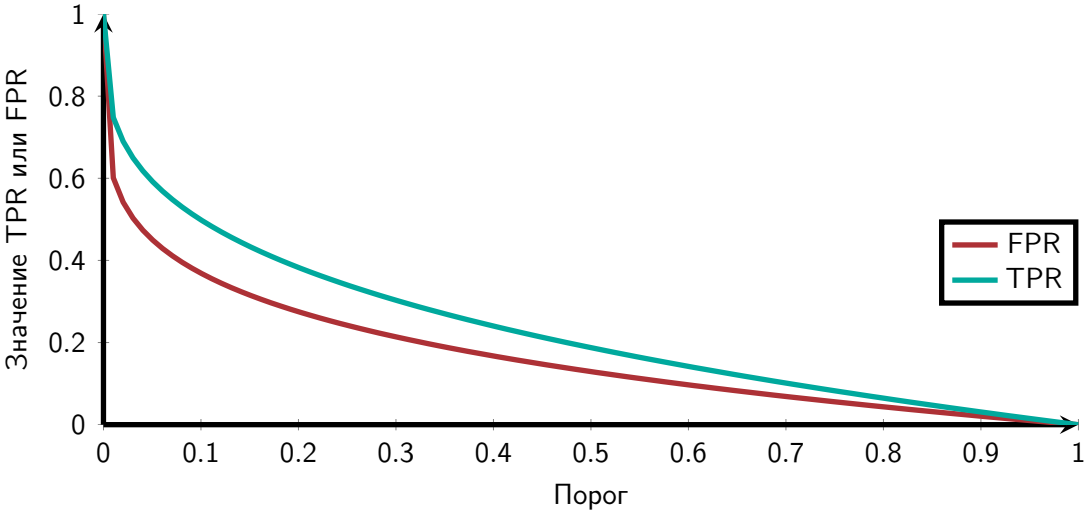
Оценивание качества прогноза

Графическая иллюстрация ROC-кривой



Оценивание качества прогноза

Графическая иллюстрация связи порога, TPR и FPR



Оценивание качества прогноза

Интерпретация ROC-кривой

- Чем выше ROC-кривая в той или иной точке, тем выше TPR при прочем равном FPR.
- Следовательно, чем выше лежит ROC-кривая, тем лучше.
- Интеграл ROC-кривой тем больше, чем выше лежит соответствующая кривая.
- Этот интеграл именуется **AUC** (area under curve) и отражает среднее качество прогноза модели при различных порогах.
- AUC принимает значения от 0 до 1, где 0.5 соответствует случайному гаданию, а 1 – идеальной модели, не совершающей ошибок при прогнозировании.
- **Преимущество** – агрегирует предиктивные способности модели при различных порогах.
- **Проблема** – отражает среднюю температуру по больнице. В частности, AUC может быть велик за счет хороших предиктивных способностей моделей при тех порогах, которые нам могут быть неинтересны исходя из содержательной экономической задачи.
- **Решение** – использовать **частный AUC** (pAUC - partial AUC), интегрируя ROC кривую на каком-то отдельном участке, например, от 0.1 до 0.3, что можно обозначить $\text{pAUC}(0.1, 0.3)$.

Оценивание гиперпараметров

Оценивание гиперпараметров с помощью кросс-валидации

- Некоторые параметры не оцениваются непосредственно при обучении модели.
- К таким параметрам, именуемым **гиперпараметрами**, можно отнести, например, количество ближайших соседей и структуру Байесовской сети.
- Подбор оптимальных значений гиперпараметров именуется **тюнингом** и, как правило, осуществляется с помощью кросс-валидации по следующему алгоритму:
  1. Выбирается метрика качества модели, например, ACC или AUC.
  2. Модель оценивается при различных значениях гиперпараметров, например, с различным числом ближайших соседей и различными метриками расстояния.
  3. Предпочтение отдается гиперпараметрам, максимизирующим выбранную метрику.
- В самом простом случае используется **жадный алгоритм**, при котором перебираются все задаваемые пользователем комбинации гиперпараметров.
- При достаточно большом числе гиперпараметров жадный алгоритм слишком ресурсозатратен, поэтому в качестве альтернативы применяются различные подходы **рандомизированного** поиска, когда гиперпараметры перебираются частично случайным образом, например, симулируются из некоторого совместного распределения.

- Представим, что в качестве гиперпараметров выступает, во-первых, число ближайших соседей $k \in \{1, 2, 3\}$, во вторых, метрика расстояния: Евклидова ($\lambda = 2$) или Манхэттан ($\lambda = 1$).
- На обучающей выборке была проведена 5-частная кросс-валидации на основании метрик ACC и F1.

k	λ	ACC	F1
1	1	0.81	0.85
2	1	0.76	0.91
3	1	0.75	0.86
1	2	0.68	0.88
2	2	0.83	0.91
3	2	0.75	0.92

- Результаты кросс-валидации свидетельствуют в пользу того, что в соответствии с ACC оптимальными являются значения $k = 2$ и $\lambda = 2$, а согласно F1 величины $k = 3$ и $\lambda = 2$.

Оценивание гиперпараметров

Валидационная выборка

Выборка	Обучающая 60%-80%	Валидационная 10%-30%	Тестовая 10%-30%
Цель	Оценивание параметров	Оценивание гиперпараметров	Проверка качества

- **Проблема** – если для тюнинга гиперпараметров использовать тестовую выборку, то они подстраиваются под нее, что может привести к подгонке гиперпараметров под тестовые данные. Это, в частности, осложняет обнаружение проблемы переобучения.
- **Решение** – оценить гиперпараметры на **валидационной выборке**, занимающей промежуточное положение между обучающей и тестовой. Качество модели с обученными на валидационной выборке гиперпараметрами проверяется на тестовой выборке.
- **Важно** - в качестве валидационной выборки обычно рассматривают части (folds), на которые обучающая выборка делится в ходе кросс-валидации (как в рассмотренном ранее примере). Однако, валидационная выборка может быть и непосредственно отделена от обучающей.

Использование прогнозов

Выигрыш (gain)

- Рассмотрим модель, прогнозирующую, купит ли индивид товар, посмотрев контекстную рекламу.
- Логично сперва показать рекламу индивидам, которые, в соответствии с оценками нашей модели, имеют **наибольшую вероятность** совершения покупки после просмотра рекламы.
- **Вопрос** – отобрав 10% клиентов с самыми большими оценками вероятностей покупки, какой процент потенциальных покупателей мы охватим?
- Например, из 1000 индивидов в тестовой выборке 200 готовы купить наш товар. Из них 50 вошли в 10% с наибольшими предсказанными вероятностями покупки. В таком случае число угаданных покупателей, то есть **выигрыш от первой децили**, обозначаемый gain(1), составит $(50/200) * 100 = 25\%$.
- В результате показав рекламу 100 наиболее вероятным покупателям из 1000 мы бы угадали 50 покупателей из 200.
- Если бы мы показывали рекламу не руководствуясь моделью, а случайным образом, то отобрали бы, вероятно, лишь $200/10 = 20$ потенциальных покупателей.
- По аналогии можно рассмотреть выигрыш от остальных децилей gain(k), где $k \in \{1, .., 10\}$.
- **Вывод** – сперва показываем рекламу наиболее вероятным покупателям и останавливаемся на децили k, если gain(k) слишком мал, то есть издержки на работу с потенциальными клиентами не покроются успешным нахождением новых клиентов.

Использование прогнозов

Пример использования выигрыша

- Допустим, что общее число посмотревших рекламу в тестовой выборке равняется 1000 и 200 из них совершили покупку.
- Предположим, что издержки на показ рекламы клиенту составляют 1 рубль, а выигрыш от продажи равняется 10 рублям.

Дециль	Купили	Выигрыш	Кумулятивный выигрыш	Выручка	Затраты	Прибыль
1	50	0.25	0.25	500	100	400
2	40	0.2	0.45	400	100	300
3	35	0.175	0.625	350	100	250
4	25	0.125	0.75	250	100	150
5	20	0.1	0.85	200	100	100
6	8	0.04	0.89	80	100	-20
7	10	0.05	0.94	100	100	0
8	2	0.01	0.95	20	100	-80
9	5	0.025	0.975	50	100	-50
10	5	0.025	1	50	100	-50

- **Вывод** – выгодно опрашивать лишь группы клиентов, входяших в первые 5 децилей.

Использование прогнозов

Подъем (lift)

- Для сравнения качества обученной модели со случайным угадыванием часто используют показатель, именуемый **подъемом** (lift).
- По аналогии с выигрышем выборка разбивается на децили в зависимости от вероятностей 1.
- Подъем для k-й децили рассчитывается как:

$$\text{lift}(k) = \frac{\text{число 1 в } k\text{-й децили}}{\text{ожидаемое число 1 в } k\text{-й децили при использовании случайного разбиения}}$$

  Где ожидаемое число 1 при случайном разбиении, очевидно, рассчитывается как отношение числа 1 во всей выборке к количеству децилей, то есть к 10.

- Вернемся к примеру с 1000 индивидами в тестовой выборке, 200 из которых готовы купить наш товар. Из них 50 вошли в 10% с набиольшими предсказанными вероятностями покупки, откуда:

$$\text{lift}(1) = \frac{50}{200/10} = 2.5$$

- Полученный результат говорит о том, что рассмотрении 1-й децили, полученной по оценкам условных вероятностей нашей модели, позволило выявить в 2.5 раза больше покупателей, чем мы бы ожидаемо выявили бы при случайном отборе индивидов (без использования модели).

Использование прогнозов

Кумулятивный подъем (cumulative lift) и пример его использования

- Кумулятивный подъем рассчитывается как:

$$\text{cumulative lift}(k) = \frac{\text{число 1 в первых } k \text{ децилях}}{\text{ожидаемое число 1 в первых } k \text{ децилях}}$$

- Пусть из 1000 посмотревших рекламу индивидов 200 совершили покупку.

Дециль	Купили	Выигрыш	Кумулятивный выигрыш	Подъем	Кумулятивный подъем
1	50	0.25	0.25	2.5	2.5
2	40	0.2	0.45	2	2.25
3	35	0.175	0.625	1.75	2.08
4	25	0.125	0.75	1.25	1.88
5	20	0.1	0.85	1	1.7
6	8	0.04	0.89	0.4	1.48
7	10	0.05	0.94	0.5	1.34
8	2	0.01	0.95	0.1	1.19
9	5	0.025	0.975	0.25	1.08
10	5	0.025	1	0.25	1

- **Вывод** – согласно кумулятивному подъему, опросив индивидов из первых 5 децилей мы ожидаемое найдем в 1.7 раза больше покупателей, чем при случайном опросе.

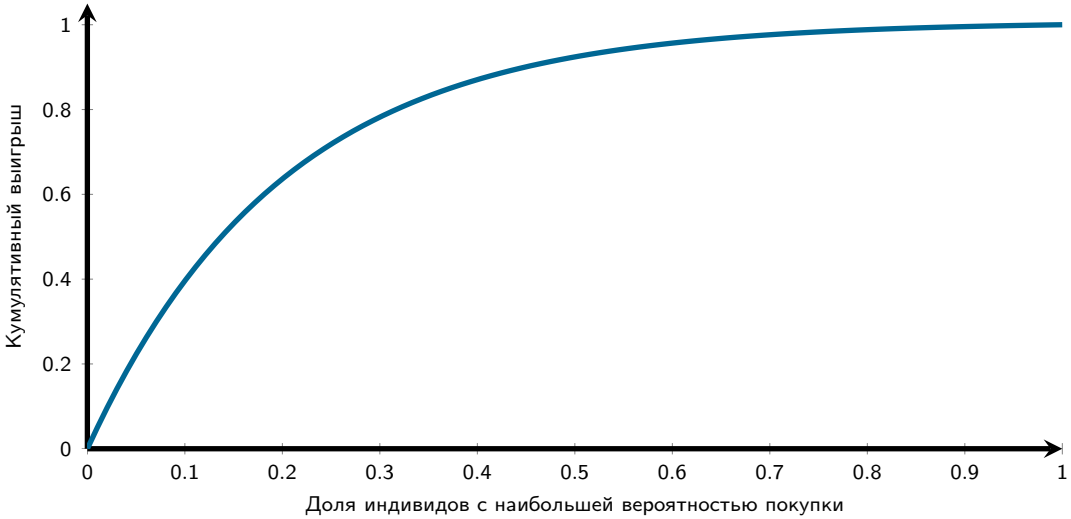
Использование прогнозов

Расчет кумулятивного выигрыша и подъема без использования децилей (gain)

- **Проблема** – при использовании децилей мы видим кумулятивные выигрыши и подъемы лишь для соответствующих децилей. Иногда мы хотим более развернутую информацию, например, о выигрыше в случае, если мы таргетируем 57% наиболее вероятных покупателей.
- **Решение** – нарисовать график, на который будут нанесены кумулятивные выигрыши и подъемы для каждой выборочной квантили наших данных:
  - Как и ранее сортируем индивидов в зависимости от вероятности покупки.
  - Для i-го индивида кумулятивный выигрыш рассчитывается как сумма его выигрыша и выигрышей всех тех, кто с большей вероятностью чем он совершит покупку.
  - Рассчитанные вероятности наносятся на график, где по оси x указана рассматриваемая доля индивидов с наибольшей вероятностью покупки, а по оси y – кумулятивный выигрыш от этих индивидов.
  - Для кумулятивного подъема и кумулятивной прибыли процедура аналогична.
- **Вывод** – для наглядности сперва можно привести графики или таблицы для децилей, а затем, для большей конкретики, рассмотреть непрерывный график.

Использование прогнозов

Графическая иллюстрация кумулятивного выигрыша



Нерепрезентативная выборка

Балансировка классов с помощью изменения структуры выборки

- Иногда выборка оказывается несбалансированной не только по причине неравномерного распределения классов в генеральной совокупности, но и из-за дизайна исследования (подхода к сбору данных).
- Например, компания, продающая телефоны, хочет научиться прогнозировать, купит ли индивид их продукт.
- На основании опроса тех, кто купил у компании телефон, были собраны данные о характеристиках 100 индивидов. Также, компания опросила 400 случайно выбранных на улице индивидов, из которых лишь 10 купили телефон, продаваемый фирмой.
- На основании данных о продажах и населения страны фирма знает, что лишь 2% индивидов покупают ее телефон.
- Объединяя исходные и собранные данные фирма получает, что в выборке имеется 500 индивидов, из которых 110 купили телефон. При этом если бы фирма опросила случайным образом на улице всех 500 индивидов, то приблизительно лишь 10 из оказались бы покупателями телефона.
- **Проблема** – модели, обученные на выборке, собранной фирмой, могут завышать условные вероятности покупки телефона случайным индивидом.
- **Undersampling** – случайным образом удаляем представителей одного из классов до тех пор, пока соотношения классов не окажутся репрезентативны генеральной совокупности. Например, из 110 индивидов, купивших телефон, оставляем лишь 10.
- **Oversampling** – случайным образом дублируем представителей одного из классов до тех пор, пока соотношения классов не окажутся репрезентативны генеральной совокупности. Например, чтобы индивиды, не купившие телефон, составили 98% от выборки, необходимо при помощи выбора с возвращением 390 превратить в x наблюдений, где x получается из решенения $x/(x + 110) = 0.98$, откуда $x = 5390$.