

Машинное обучение в экономике

Метод ближайших соседей

Потанин Богдан Станиславович

доцент, научный сотрудник, кандидат экономических наук

2023–2024

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.

Метод ближайших соседей

Метрика расстояния

- 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:

- Минковский:** $d(x, y) = \left(\sum_{i=1}^m (x_i - y_i)^\lambda \right)^{\frac{1}{\lambda}}$, где $\lambda \in \mathbb{N}$

Чем больше λ , тем больше штраф за разницу 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: $x = (4, 2, 8)$, $y = (0, 5, 8)$, $d(x, y) = |4-0| + |2-5| + |8-8| = 7$

Метод ближайших соседей

Случай с единственным соседом

- **Идея** – предсказываем наблюдению значение, совпадающее со значением другого, наиболее близкого к нему наблюдения, именуемого **ближайшим соседом**.
- Через X_i обозначим i -е наблюдение и запишем классифицирующее правило:

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

- Прогноз $\hat{y}(x)$ это Y_i ближайшего соседа x с точки зрения функции расстояния $d()$.
- 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 дистанции Евклида:

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

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

Метод ближайших соседей

Несколько соседей

- Прогноз, основанный всего на одном наблюдении, будет обладать очень большой дисперсией, что мотивирует учет большего числа соседей.
- По аналогии с одномерным случаем находим k ближайших соседей и выбираем наиболее часто встречающееся среди них значение целевой переменной.
- Число соседей k 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:

$$d(x, X_1) = |1 - 0| + |3 - 1| = 3 \quad d(x, X_2) = |1 - 1| + |3 - 2| = 1$$

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

- 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

Взвешенные ближайшие соседи

- **Проблема** – при малом числе соседей k метод ближайших соседей страдает от большой дисперсии, а при большом количестве соседей – склонен к переобучению.
- **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 $q_1(x), \dots, q_k(x)$ признаки ближайших соседей наблюдения x , а через $q_1(y), \dots, q_k(y)$ – значения их целевых переменных.
- **Интуиция** – чем больше расстояние $d(x, q_i(x))$, тем меньший вес должен получить i -й сосед. Этой идее удовлетворяет, например, функция $1/d(x, q_i(x))^2$.
- Классификатор присваивает значение 1, если суммарный вес соседей с $q_i(y) = 1$ больше, чем суммарный вес соседей с $q_i(y) = 0$:

$$\hat{y}(x) = I \left(\sum_{i: q_i(y)=1} \frac{1}{d(q_i(x), x)^2} \geq \sum_{i: q_i(y)=0} \frac{1}{d(q_i(x), x)^2} \right)$$

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, измеряном 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 $x = (25, 180)$, $X_1 = (50, 190)$ 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 максимальное значение и поделив на разницу между максимальным и минимальным значениями.

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):

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

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

- 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**:

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

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

Оценивание качества прогноза

Несбалансированная выборка

- 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 $\text{recall}(1)$ и $\text{recall}(0)$ полноту, посчитанную для 1 и 0 соответственно. В последнем случае формула будет иметь вид $\text{recall}(0) = \frac{TN}{TN+FP}$.
- **Средняя точность по классам** рассчитывается как:

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

- Достигнуть большой средней точности по классам можно и в случае, когда лишь один из $\text{recall}(1)$ или $\text{recall}(0)$, а другой – гораздо меньше. Для того, чтобы повысить штраф за крайне малое значение одного из этих показателей среднюю точность по классам часть считают с помощью гармонического среднего:

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

Оценивание качества прогноза

ROC-кривая

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

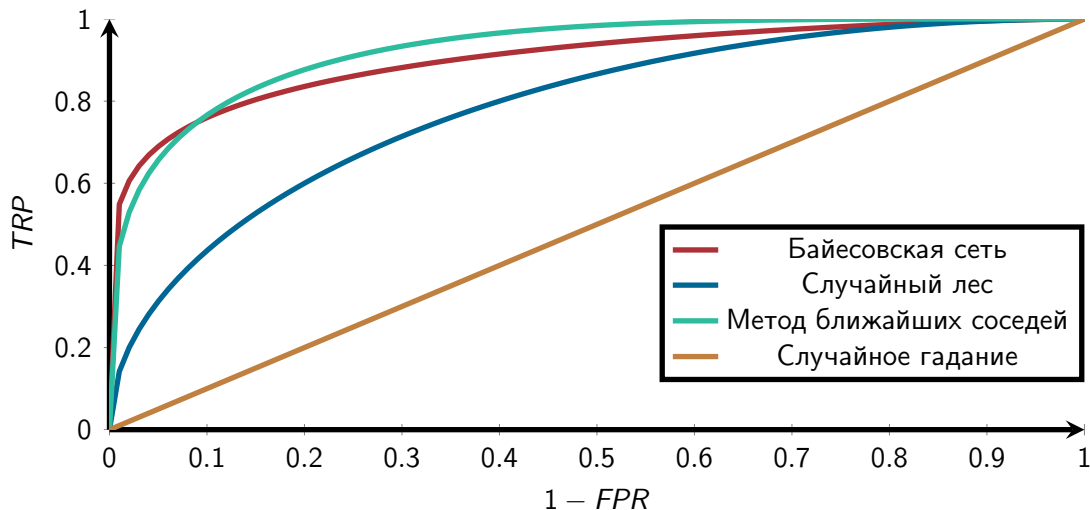
$$TPR = \frac{TP}{TP + FN} \quad \text{доля предсказанных 1 среди всех истинных 1}$$

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

- **Важно** – по мере уменьшения порога c растут FPR (false positive rate - плохо) и TPR (true positive rate - хорошо).
- Перебирая различные значения порога (отсортированные оценки вероятностей) мы получаем все возможные комбинации TRP и FPR в наших данных.
- **ROC-кривая** отражает график зависимости между TPR (ось x) и $1 - FPR$ (ось y).
- 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 TRP , а значит эта модель предпочтительна при соответствующем пороге.
- **Применение** – можно понять, при каких порогах различные модели имеют преимущество.

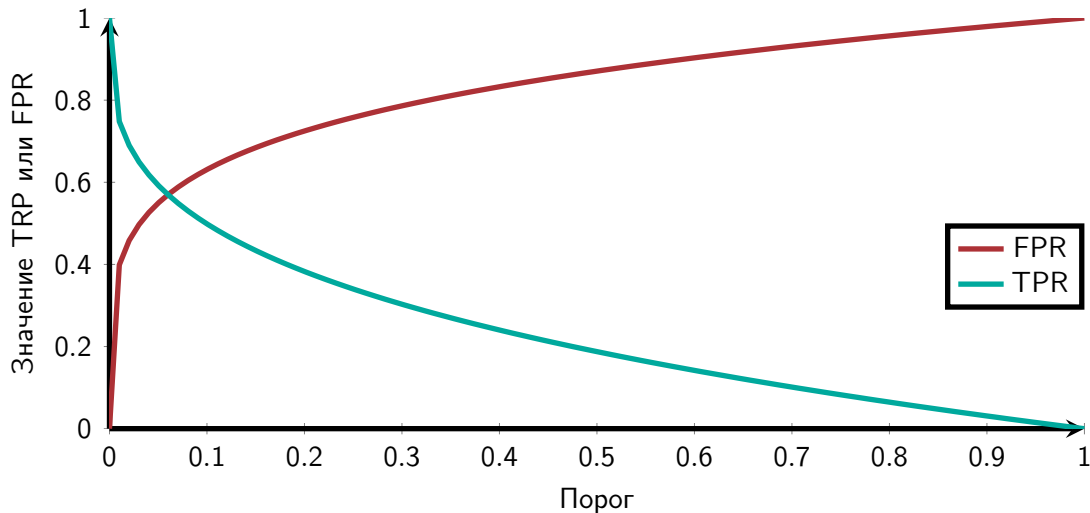
Оценивание качества прогноза

Графическая иллюстрация ROC-кривой



Оценивание качества прогноза

Графическая иллюстрация связи порога, TRP и FPR



Оценивание гиперпараметров

Оценивание гиперпараметров с помощью кросс-валидации

- Чем выше ROC-кривая в той или иной точке, тем выше TRP при прочем равном $1 - FPR$.
- Следовательно, чем обычно 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.
- **Преимущество** – агрегирует предиктивные способности модели при различных порогах.
- **Недостаток** – 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 содержательное экономической задачи.

Оценивание качества прогноза

Выигрыш (gain)

- Рассмотрим модель, предсказывающую, купит ли индивид товар, посмотрев контекстную рекламу.
- Логично сперва показать рекламу индивидам, которые, в соответствии с оценками нашей модели, имеют **наибольшую вероятность** совершения покупки.
- **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 $\text{gain}(1)$, 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, случайным образом, то отобрали бы, вероятно, лишь $200/10 = 20$ потенциальных покупателей.
- По аналогии можно рассмотреть выигрыш от остальных децилей $\text{gain}(k)$, где $k \in \{1, \dots, 10\}$.
- **Вывод** – сперва показываем рекламу наиболее вероятным покупателям и останавливаемся на децили k , если $\text{gain}(k)$ слишком мал, то есть издержки на работу с потенциальными клиентами не покроются успешным нахождением новых клиентов.

Оценивание качества прогноза

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 рублей, 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 | 0.65 | 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, входящих в первые 5 децилей.

Оценивание гиперпараметров

Оценивание гиперпараметров с помощью кросс-валидации

- Некоторые параметры не оцениваются непосредственно при обучении модели.
- К таким параметром, именуемым **гиперпараметрами**, можно отнести, например, глубину решающего дерева и количество ближайших соседей.
- Подбор оптимальных значений гиперпараметров именуется **тюнингом** и, как правило, осуществляется с помощью кросс-валидации по следующему алгоритму:
 - ❶ Выбирается метрика качества модели, например, точность АСС прогноза или взвешенные доли прогнозов различного вида.
 - ❷ Модель оценивается при различных значениях гиперпараметров, например, с различным числом ближайших соседей и различными метриками расстояния.
 - ❸ 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.

Оценивание гиперпараметров

Пример с жадным алгоритмом

- 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 ($\lambda = 1$).
- Выборка была разбита на обучающую и тестовую, после чего была проведена 5-блочная 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 $\lambda = 2$.