Машинное обучение в экономике

Байесовские сети

Потанин Богдан Станиславович

доцент, научный сотрудник, кандидат экономических наук

2023–2024

- Методы классификации:
  - Наивный байесовский классификатор.
  - Байесовский классификатор.
  - Байесовские сети.
- Базовые понятия:
  - Проклятье размерности и фрагментация данных.
  - Переобучение.
  - Точность прогноза.
  - Обучающая и тестовая выборки.
  - Кросс-валидация.
  - Графические модели.

Байесовский классификатор

Основная идея

- Предположим, что бизнес заинтересован в нахождении клиентов, готовых совершить покупку.
- У бизнеса имеется информация о характеристиках клиентов, на основании которой необходимо определить потенциальных покупателей.
- **Вопрос** – купит ли товар женщина средних лет с высшим образованием?
- **Интуитивный ответ** – посмотрим на долю женщин средних лет с высшим образованием, которые раньше купили товар. Эта доля будет отражать вероятность покупки товара такой женщиной. Если соответствующая доля достаточно велика (например, больше 0.5), то будем прогнозировать совершение покупки.
- Например, если из 100 женщин среднего возраста с высшим образованием 80 совершили покупку, то вероятность покупки для этой категории покупателей оенивается как 0.8 и разумно прогнозировать, что такой клиент купит товар.

Байесовский классификатор

Оценивание условных вероятностей

- Имеется n-мерный вектор значений **целевой переменной** Y и матрица **признаков** X с n строками (число наблюдений) и m столбцами (количество признаков).
- Для простоты допустим, что целевая переменная и признаки являются бинарными переменными $Y_i \sim \text{Ber}(p)$, $X_{ij} \sim \text{Ber}(p_j)$, где $i \in \{1, ..., n\}$ и $j \in \{1, ..., m\}$, а также $p, p_j \in (0, 1)$.
- Из базового курса статистики известно, что состоятельная оценка $\hat{p}$ может быть получена как:

$$\hat{p} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad \textbf{(доля единиц в векторе Y)}$$

- Рассмотрим оценку условной вероятности $p_x = P(Y_i = 1 | X_i = x_i)$ того, что целевая переменная примет значение 1, при условии, что вектор признаков оказался равен $x_i \in \{0, 1\}^m$.
- Обозначим через $n(x)$ число наблюдений таких, что $X_i = x$, а через $n(x, y)$ число наблюдений таких, что $X_i = x$ и $Y_i = 1$.

$$n(x) = |\{i \in \{1, ..., n\} : X_i = x\}| \qquad n(x, y) = |\{i \in \{1, ..., n\} : Y_i = 1 \wedge X_i = x\}|$$

  Где $|A|$ ообозначет число элементов множества A.
- Поскольку $n(x, y)/n$ – это состоятельная оценка $P(X_i = x, Y_i = 1)$ и $n(x)/n$ – это состоятельная оценка $P(X_i)$, то по теореме Слуцкого состоятельная оценка p_x может быть получена как $\hat{p}_x = \frac{n(x,y)}{n(x)}$.

- Введем функцию индикатор:

$$I(\text{условие}) = \begin{cases} 1, \text{ если условие соблюдено} \\ 0, \text{ в противном случае} \end{cases}$$

- Например, $I(5 = 5) = 1$, $I(5 = 3) = 0$ и $I(5 \geq 3) = 1$.
- **Бинарный классификатор** определяет значение бинарной целевой переменной $y \in \{0, 1\}$ в зависимости от значения вектора признаков $x \in \{0, 1\}^m$, например, как:

$$\hat{y}(x) = \begin{cases} 1, \text{ если } \hat{p}_x \geq c \\ 0, \text{ в противом случае} \end{cases}$$

- Порог c может варьироваться в зависимости от задачи. Для простоты пока что остановимся на случае $c = 0.5$, поскольку при $\hat{p}_x \geq 0.5$ мы оцениваем вероятность 1 как большую, чем вероятность 0.

- Рассмотрим задачу удержания клиента. Владелец мобильного приложения хочет определить пользователей, которые могут прекратить использовать приложение, чтобы заблаговременно предпринять меры по их удержанию. Для этого необходимо научиться прогнозировать таких пользователей.

Переменные

Y – прекратил ли клиент пользоваться приложением в прошлом месяце (1 – да, 0 – нет).
X_{*1} – сократилась ли частота использования приложения в позапрошлом месяце (1 – да, 0 – нет).
X_{*2} – изменились ли основные функции, используемые пользователем в позрапрошлом месяце (1 – да, 0 – нет).

Данные

Y	X_{*1}	X_{*2}
0	1	0
1	1	1
1	1	0
1	0	1
0	1	1
1	0	1
0	0	0
1	1	0

- Пусть $x = (1, 0)$, тогда $n(x) = 3$ и $n(x, y) = 2$, откуда $\hat{p}_x = 2/3$.
- Поскольку $\hat{y}(x) = I(\hat{p}_x > 0.5) = 1$, то модель говорит о том, что клиент прекратит использовать приложение, а значит можно попытаться удержать его какими-то бонусным предложением.

Байесовский классификатор

Идеи проклятья размерности и фрагментации данных

- Вернемся к примеру с бизнесом, заинтересованном в поиске клиентов, готовых совершить покупку.
- Допустим, что у бизнеса имеется очень большое количество информации о клиентах.
- **Вопрос** – с какой вероятностью совершит покупка женщина средних лет с высшим образованием, **тремя детьми и ипотекой**.
- Из 100 женщин с высшим образованием и тремя детьми лишь 10 имеют трех детей и ипотеку. Из них 7 совершили покупку, откуда оценка вероятности окажется 0.7.
- **Первая проблема** – оценка вероятности, полученная по столь малому числу наблюдений, неэффективна. Мы не можем надежно судить о поведении клиентов с такими характеристиками лишь по 10 наблюдениям.
- **Вторая проблема** – если мы учтем дополнительные признаки, например, наличие машины и работы, то в данных может вовсе не оказаться наблюдений по таким клиентам, то есть удовлетворяющим сразу всем признакам.
- Обе проблемы отражают **проклятье размерности** и возникают из-за того, что при использовании большого числа признаков наша выборка разбивается на большое число подвыборов, что именуется **фрагментацией данных**. При этом в некоторых из этих подвыборок недостает наблюдений для того, чтобы оценить интересующие нас характеристики.
- Например, у нас слишком мало наблюдений по женщинам средних лет с высшим образованием, тремя детьми и ипотекой (подвыборка), чтобы оценить для них вероятность покупки (характеристика распределения).

- Эффективность оценки p_x зависит от $n(x, y)$.
- На практике число признаков m часто оказывается достаточно велико. В таком случае $n(x)$ и $n(x, y)$ при большинстве x оказываются достаточно малыми числами.
- Например, в выборке может быть очень мало бабушек (признак 1) с высшим образованием (признак 2), (много других признаков), у которой внук полетел в космос (признак 1000).
- Данная проблема именуется **проклятием размерности** и возникает в байесовском классификаторе из-за того, что число оцениваемых параметров совпадает с числом возможных условных вероятностей, которое достаточно быстро растет вместе с количеством признаков m и совпадает с числом возможных значений x, которое равно 2^m, что приводит к стремительному росту дисперсии оценок вследствие **фрагментации данных**.
- **Идея решения проблемы** – сократить число параметров, за счет введения ограничений на условные вероятности.

Наивный байесовский классификатор

Идея условной независимости

- Мы пытаемся предсказать, купит ли нашу компьютерную игру индивид, в зависимости от его пола и наличия дорогой игровой видеокарты.
- **Предположим**, что в целом женщины реже увлекаются компьютерными играми, а значит реже будут покупать дорогие видеокарты.

$$P(\text{Дорогая видеокарта}|\text{Женщина}) < P(\text{Дорогая видеокарта}|\text{Мужчина})$$

- Однако, среди тех, кто увлекается компьютерными играми, наличие дорогой видеокарты и пола может быть не связано.
- Поскольку нашу игру покупают те, кто увлекаются компьютерными играми, разумно предположить, что среди купивших игру пол и наличие дорогих видеокарт распределены независимо:

$$P(\text{Дорогая видеокарта}|\text{Женщина}, \text{Купил игру}) = P(\text{Дорогая видеокарта}|\text{Мужчина}, \text{Купил игру})$$

- Таким образом, наличие дорогой видеокарты и пол **условно независимы** при условии покупки игры:

$$P(\text{Женщина, дорогая видеокарта}|\text{Купил игру}) =$$
$$= P(\text{Дорогая видеокарта}|\text{Купил игру}) \times P(\text{Женщина}|\text{Купил игру})$$

Наивный байесовский классификатор

Идея применения формулы Байеса при условной независимости

- Запишем вероятность покупки игры с помощью формулы Байеса, учитывая условную независимость:

$$P(\text{Купил игру}|\text{Женщина с дорогой видеокартой}) = \quad \textbf{формула Байеса}$$

$$= \frac{P(\text{Купил игру}) \times P(\text{Женщина, дорогая видеокарта}|\text{Купил игру})}{P(\text{Женщина, дорогая видеокарта})} = \quad \textbf{условная независимость}$$

$$= \frac{P(\text{Купил игру}) \times P(\text{Дорогая видеокарта}|\text{Купил игру}) \times P(\text{Женщина}|\text{Купил игру})}{P(\text{Женщина, дорогая видеокарта})}$$

- Поскольку женщины редко покупают дорогие видеокарты, то в выборке может быть очень мало наблюдений по женщинам с дорогими видеокартами, что осложняет оценивание **совместной** условной вероятности:

$$P(\text{Женщина, дорогая видеокарта}|\text{Купил игру})$$

- Однако в выборке может быть много информации по отдельности о покупках игры среди женщин и среди тех, у кого есть дорогая видеокарта. То есть легко оценить **маржинальные** условные вероятности.

$$P(\text{Дорогая видеокарта}|\text{Купил игру}) \qquad P(\text{Женщина}|\text{Купил игру})$$

- **Ключевая идея** – неэффективно оцениваемая из-за фрагментации данных совместная условная вероятность благодаря **предположению об условной независимости** записывается как произведение эффективно оцениваемых маржинальных условных вероятностей.

- Необходимо избавиться от проклятья размерности не только в числителе, но и в знаменателе условной вероятности, для чего применим формулу полной вероятности:

$$P(\text{Женщина, дорогая видеокарта}) =$$
$$P(\text{Женщина, дорогая видеокарта}|\text{Купил игру})P(\text{Купил игру}) +$$
$$+P(\text{Женщина, дорогая видеокарта}|\text{Не купил игру})P(\text{Не купил игру}) =$$
$$= P(\text{Женщина}|\text{Купил игру})P(\text{Дорогая видеокарта}|\text{Купил игру})P(\text{Купил игру}) +$$
$$P(\text{Женщина}|\text{Не купил игру})P(\text{Дорогая видеокарта}|\text{Не купил игру})P(\text{Не купил игру})$$

- Мы вновь расписали совместную вероятность через маржинальные, оценивание которых возможно, как правило, по большому числу наблюдений.
- Рассчитав знаменатель соответствующим образом мы можем предсказывать покупку игры в случаях, когда вероятность покупки больше, чем отсутствия покупки, что выполняется в случае:

$$P(\text{Купил игру}|\text{Женщина, дорогая видеокарта}) > 0.5$$

- Носителем дискретной случайной величины X называется множество значений, которое случайная величина принимает с ненулевой вероятностью:

$$\text{supp}(X) = \{x \in R : P(X = x) > 0\}$$

- Формула полной вероятности:

$$P(X = x) = \sum_{y \in \textbf{supp}(Y)} P(X = x|Y = y) \times P(Y = y)$$

- Формула Байеса:

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{\sum_{t \in \textbf{supp}(Y)} P(X = x|Y = t)P(Y = t)}, \qquad \text{где } x \in \text{supp(X)}$$

- Дискретные случайные величины $X_1, X_2, ..., X_n$ условно независимы при условии случайной величины Y, если при любых $x_i \in \text{supp}(X_i)$ и $y \in \text{supp}(Y)$, где $i \in \{1, ..., n\}$, выполняется:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n|Y = y) = \prod_{i=1}^{n} P(X_i = x_i|Y = y)$$

- Предпожим, что признаки X_{ij} условно независимы при условии Y_i, и применим формулу Байеса:

$$p_x = P(Y_i = 1|X_i = x) = \frac{P(Y_i = 1)P(X_i = x|Y_i = 1)}{P(X_i = x|Y_i = 1)P(Y_i = 1) + P(X_i = x|Y_i = 0)P(Y_i = 0)} =$$
$$= \frac{P(Y_i = 1)\prod_{j=1}^{m} P(X_{ij} = x_j|Y_i = 1)}{P(Y_i = 1)\prod_{j=1}^{m} P(X_{ij} = x_j|Y_i = 1) + P(Y_i = 0)\prod_{j=1}^{m} P(X_{ij} = x_j|Y_i = 0)}$$

- Вероятности $P(Y_i = t)$ именуются **априорными** и оцениваются как доля Y_i таких, что $Y_i = t$.
- Вероятности $P(X_{ij} = x_j|Y_i = t)$ называются **факторами** и оцениваются как доля случаев $X_{ij} = x_j$ среди наблюдений, у которых $Y_i = t$.
- При оценивании факторов мы не сталкиваемся с проклятьем размерности, поскольку считаем долю встречающихся значений лишь для одного признака, а не сразу для всех.
- Мы описываем 2^{m+1} совместные условные вероятности $P(X_i = x|Y_i = t)$ с помощью $4m$ факторов $P(X_{ij} = x_j|Y_i = t)$, что и позволяет избежать проклять размерности за счет того, что на каждый оцениваемый параметр приходится больше наблюдений.
- Поскольку у p_x и $1 - p_x$ одинаковый знаменатель, то для удобства **классифицирующее правило** наивного Байесовского классификатора иногда формулируют как $\hat{y}(x) = I(\hat{p}_x/(1 - \hat{p}_x) > 1)$.

- Рассмотрим задачу предсказания дефолта заемщика.

Переменные

Y – дефолт по кредиту (1 – случился, 0 – не случился).
X_{*1} – наличие высшего образования (1 – есть, 0 – нет).
X_{*2} – наличие детей (1 – есть, 0 – нет).

Данные

Y	X_{*1}	X_{*2}
1	1	0
0	0	1
0	1	0
1	1	0
1	0	1

- Пусть $x = (1, 0)$, тогда Байесовский классификатор выдаст $\hat{y}(x) = I\left(\frac{2}{3} \geq 0.5\right) = 1$.
- Наивный Байесовский классификатор в данном случае даст идентичный ответ:

$$\hat{y}(x) = I\left(\frac{\hat{P}(Y_i = 1)\hat{P}(X_{1i} = 1|Y_i = 1)\hat{P}(X_{2i} = 0|Y_i = 1)}{\hat{P}(Y_i = 0)\hat{P}(X_{1i} = 1|Y_i = 0)\hat{P}(X_{2i} = 0|Y_i = 0)} \geq 1\right) = I\left(\frac{\frac{3}{5} \times \frac{2}{3} \times \frac{1}{3}}{\frac{2}{5} \times \frac{1}{2} \times \frac{1}{2}} \geq 1\right) = I\left(\frac{4}{3} \geq 1\right) = 1$$

- Обратитим внимание, что обычный Байесовский классификатор не позволяет получить прогноз для $x = (1, 1)$, поскольку такой комбинации признаков нет в данных, а наивный – позволяет.

Оценивание качества модели

Доля верных прогнозов и разделение выборки на обучающую и тестовую

- Обозначим через $\hat{Y}_i = \hat{y}(X_i)$ прогнозы нашего классификатора и оценим **точность прогноза** как долю верных прогнозов в нашей выборке:

$$\text{ACC} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = \hat{Y}_i)$$

- **Проблема** – даже если наша модель выдает точные прогнозы на той же выборке, на которой она оценивалась, это не значит, что она будет хорошо работать на новых данных.
- **Решение** – разделить выборку на **обучающую** и **тестовую**.
- **Обучающая выборка** – выборка, на которой оцениваются параметры модели.
- **Тестовая выборка** – выборка, не входящая в обучающую выборку, используемая лишь для оценивания точность обученной модели.
- Например, из 1000 наблюдений по дефолтам на 800 мы можем обучить Наивный Байесовский классификатор, а на оставшихся 200 – посчитать точность прогноза ACC.

Оценивание качества модели

Переобучение модели

- Модель **переобучилась**, если точность прогнозов на обучающей выборке гораздо выше, чем на тестовой.
- **Аналогия** – студент может выучить наизусть решение задач из минимума по теории вероятностей (заранее известные задачи, гарантированно встречающиеся на контрольной работе лишь с другими числами) и написать его на отличную оценку (точность по тренировочной выборке), однако завалить максимум (точность по тестовой выборке), если при решении минимума студент заучивал решения задач, а не пытался разобраться в основных принципах и формулах теории вероятностей, благодаря которым были получены решения.
- **Вывод** – при изучении минимума нужно не просто на память заучивать решение задач (переобучаться), а разбираться в формулах, теоремах и основных закономерностях решения задач (обучение на тренировочной выборке), чтобы применять эти навыки для решения новых задач, отличающихся от тех, что были в минимуме (прогнозирование на тестовой выборке).
- **Мораль** – хорошая модель машинного обучения должна изучить не сами данные, а кроющиеся в них закономерности, чтобы использовать их для прогнозирования на новых, неизвестных ей на момент обучения данных.

Оценивание качества модели

Переобучение в контексте Байесовского классификатора и Наивного байесовского классификатора

- Какая модель в большей степени склонна к переобучению, Байесовский классификатор или Наивный Байесовский классификатор?
- Байесовский классификатор в большей степени склонен к переобучению, поскольку тщательно запоминает выборку за счет большого числа параметров (совместных условных вероятностей).
- Наивный Байесовский классификатор не склонен к переобучению, так как запоминает лишь информацию о маржинальных условных вероятностях (факторах) и реконструирует на ее основе информацию о совместных условных вероятностях, тем самым улавливая закономерности в данных.
- **Аналогия** – Байесовский классификатор подобен студенту, заучивающему наизусть решения всех разобраных задач. Для того, чтобы такой подход позволил хорошо написать контрольную работу (высокая точность модели), необходимо заучить очень много задач (большое число наблюдений).
- **Аналогия** – Наивный Байесовский классификатор подобен студенту, который хорошо решает задачи, требующие вдмумчивого применения отдельных формул, например (оценивание маржинальных условных вероятностей). Однако, испытывает трудности при решении сложных задач, когда эти формулы необходимо нетривиальным образом комбинировать (нарушение допущения об условной независимости). В результате нарешивание простых задач даже в большом количестве (много наблюдений) не позволит преодолеть планку в умении решать более сложные задачи (асимптотическое смещение из-за нарушения допущения об условной независимости).

- При разбиении выборки на тренировочную и тестовую части всегда есть риск, что тренировочная часть окажется гораздо проще или сложней, чем тестовая, что не позволит адекватным образом оценить точность прогнозов.
- В качестве альтернативы использованию одной тестовой выборки применяется k-**частная кросс-валидация** (k-fold cross-validation), при которой выборку разбивают на k непересекающихся равных частей.
- Каждая из этих k частей поочередно выступает в качестве тестовой выборки, для которой рассчитывается точность ACC_j, где $j \in \{1, ..., k\}$.
- Итоговая точность рассчитывается по результатам усредненяя точностей, полученных на тестовых выборках:

$$\text{ACC} = \frac{1}{k} \sum_{j=1}^{k} \text{ACC}_j$$

- Обычно полагают $k = 1$, $k = 5$ или $k = 10$. Также популярен случай $k = n$, когда в качестве тестовой выборки поочередно выступает каждое из наблюдений исходной выборки (leave-one-out-cross-validation).

- Допущение об условной независимости в наивном Байесовском классификаторе является весьма нереалистичным для большинства встречающихся на практике задач.
- Например, в задаче предсказания дефолта заемщика образование и доход будут зависимыми и для тех, у кого наступил дефолт, и для тех, у кого он не наступил.
- Байесовские сети пытаются предложить золотую середину между Байесовским классификатором и наивным байесовским классификатором, предполагая менее жесткие и более теоретически обоснованные формы независимости между переменными модели.

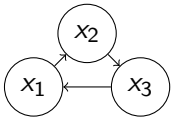


Рис.: Граф с циклом

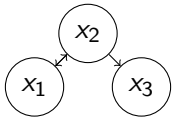


Рис.: Граф с двунаправленным ребром

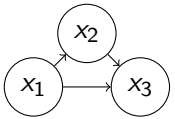


Рис.: DAG

- Ориентированные ациклические графы (DAG) состоят из узлов (nodes) и ориентированных (directed) ребер (edges).
- На графиках кружки отражают узлы, линии представляют ребра, а стрелочки отвечают за направление ребер.
- На рисунке, на котором изображен пример *DAG*, из узла x_1 выходит ребро в направлении x_2. В таком случае x_2 является **ребенком** x_1, а x_1 называется **родителем** x_2. По аналогии x_1 и x_2 являются родителями x_3.

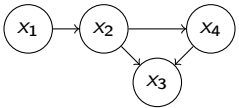


Рис.: DAG

Узлы	Факторы
X_1	$P(X_1 = x_1)$
X_2	$P(X_2 = x_2 \mid X_1 = x_1)$
X_3	$P(X_3 = x_3 \mid X_2 = x_2, X_4 = x_4)$
X_4	$P(X_4 = x_4 \mid X_2 = x_2)$

Таблица: Факторизация совместного распределения

- Байесовская сеть описывается через DAG, в котором каждому узлу соответствует случайная величина и условное распределение этой случайной величины при условии ее родителей.
- Обозначим через $\text{Parents}(X_i)$ и $\text{Children}(X_i)$ множество родителей и детей X_i.
- Байесовская сеть основана на допущении о том, что **совместные вероятности могут быть представлены как функции от факторов** следующим образом (для краткости опустим x_i):

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i))$$

Байесовские сети

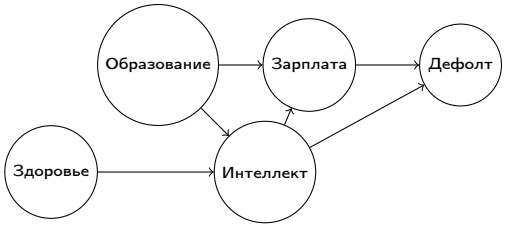
Факторизация

- Можно показать, что в Байесовской сети условные вероятности пропорциональны:

$$P(X_i|X_1, ..., X_{i-1}, X_{i+1}, ..., X_n) \propto P\left(X_i|\text{Parents}(X_i)\right) \prod_{Z \in \text{Children}(X_i)} P\left(Z|\text{Parents}(Z)\right)$$

- Поскольку у всех усовных вероятностей одинаковый знаменатель, то применяя тот же прием, что и в Байесовских сетях, нетрудно получить выражение для классификатора.
- Обычно факторы оцениваются по аналогии с обычным байесовским подходом. Например, $P(X_5 = 1|X_2 = 1, X_6 = 0)$ оценивается как доля наблюдений, у которых $X_5 = 1$, среди наблюдений, у которых $X_2 = 1$ и $X_6 = 0$.
- До тех пор, пока число родителей у каждого из факторов не слишком велико, Байесовская сеть позволяет избежать проклятья размерности, при этом накладывая менее жесткие ограничения, чем наивный Байесовский классификатор.
- Наивный Байесовский классификатор является частным случаем Байесовской сети, в которой целевая переменная выступает в качестве единственного родителя для всех признаков, не связанных между собой ребрами.

- Узел a называется **наследником** узла b, если из узла a, следуя по навлениям ребер (по стрелочкам), можно прийти в узел b. То есть наследниками узла a являются все его дети и дети этих детей.
- В Байесовской сети соблюдается **локальное марковское свойство**, а именно $X_i|\text{Parents}(X_i)$ не зависит от тех, кто **не являются** его наследниками X_i.
- Это свойство мотивирует строить DAG исходя из содержательных соображений по поводу того, что скорее является причиной (родители), а что – следствием (дети). То есть зафиксировав все причины (родителей) мы не можем уточнить вероятность события за счет уточнения информации о тех, кто не являются его наследниками (следствиями).



Например, в соответствии с графом при фиксированных образовании и интеллекте зарплата и здоровье независимы.

- Структура DAG может задаваться как на основании экспертного мненеия, так и по результатам обучения на данных.
- **Наивное решение –** (exhaustive search) обучить структуру, перебрав все возможные способы построения графа и выбрать тот, что дает наилучший результат, например, на основании вневыборочного прогноза или байесовского информационного критерия BIC.
- **Недостаток наивного решения** – даже при среднем числе признаков количество возможных графов слишком велико, чтобы перебрать их все за разумный промежуток времени.
- **Популярная альтернатива** (hill climb search) – взять за основу DAG, подобранный на основании экспертного мнения. Перебрать все DAG, отличающиеся от исходного лишь одним ребром (стрелочкой). Выбрать из них этих DAG самый лучший, например, по BIC. Повторять до тех пор, пока можно улучшить качество модели.
- **Недостаток альтернативы** – находит лишь локальный максимум и поэтому чувствителен к выбору начального DAG.
- Для уменьшения числа рассматриваемых DAG часто используют тесты (обычно хи-квадрат) на условную независимость признаков. В результате перебираются лишь те DAG, структура которых не противоречит результатам тестов.

Байесовские сети

Типичный алгоритм применения

Из содержательных соображений строится (рисуется) DAG

↓

Оцениваются факторы – условные на родителей вероятности переменных

↓

С помощью оценок факторов оценивается условная вероятность целевой переменной

↓

Применяется классификатор для прогнозирования значений целевой переменной

↓

В случае необходимости улушается структура DAG, например, с помощью hill climb search

↓

Вновь оцениваются факторы и условные вероятности, с помощью которых строются прогнозы целевой переменной

Особенности применения наивного Байесовского классификатора

Небинарная целевая переменная

- Предположим, что целевая переменная не является бинарной $\text{supp}\,(Y_i) \in \{0, ..., k\}$ и может принимать одно из $k + 1$ возможных значений.
- Например, клиент может установить бесплатную $Y_i = 0$, базовую $Y_i = 1$ или премиальную $Y_i = 2$ версию приложения.
- По аналогии с бинарным случаем вероятность $Y_i = t$ оценивается как:

$$\hat{P}(Y_i = t|X_{ij} = x_j) = \frac{\hat{P}(Y_i = t)\prod_{j=1}^{m}\hat{P}(X_{ij} = x_j|Y_i = t)}{\sum_{q=0}^{k}\hat{P}(Y_i = q)\prod_{j=1}^{m}\hat{P}(X_{ij} = x_j|Y_i = q)}$$

- Предсказывается значение t, которому соответствует наибольшая оценка условной вероятности. Поскольку знаменатель оценки условной вероятности не зависит от t, самой большой будет вероятность с наибольшим числителем, откуда получаем классифицирующее правило:

$$\hat{y}(x) = \underset{t\in\{0,\ldots,k\}}{\text{argmax}}\ \hat{P}(Y_i = t)\prod_{j=1}^{m}\hat{P}(X_{ij} = x_j|Y_i = t)$$

Особенности применения наивного Байесовского классификатора

Небинарные признаки

- Предположим, что признаки не являются бинарными переменными $\text{supp}\,(X_{ij}) \in \{0, ..., k_j\}$, то есть j-й признак может принимать одно из $k_j + 1$ возможных значений.
- Например, в качестве одного из признаков мы можем рассматривать уровень образования индивида: начальное $X_{ij} = 0$, среднее специальное $X_{ij} = 1$ и высшее $X_{ij} = 2$.
- В таком случае изменяется лишь способ оценивания условных вероятностей:

$$\hat{P}(X_{ij} = x_j | Y_i = t) = (\text{доля } X_{ij} = x_j \text{ среди } Y_i = t)$$

- Если число значений, принимаемых X_{ij} велико, то может возникнуть проблема фрагментации данных. Особенно, если число значений, принимаемых Y_i, также велико.
- Для избежания фрагментации данных можно объединить некоторые значения для признаков или целевых переменных.
- Например, если X_{ij} отражает профессию индивида и соответствующая переменная может принимать $(k_j + 1) = 100$ значений, то разумно может быть объединить некоторые из них, например, разделив все профессии на технические, полутехнические и нетехнические, перейдя к $k_j + 1 = 3$.

Особенности применения наивного Байесовского классификатора

Альтернативные формулы оценивания условных вероятностей

- Рассмотрим подробнее условную вероятность $P(X_{ij} = x_j | Y_i = t)$.
- Ранее мы оценивали эту вероятность как долю наблюдений $X_{ij} = x_j$ среди $Y_i = t$.
- В качестве альтернативы, не требующей объединения различных значений признаков, можно предположить конкретную форму условного распределения.
- Например, можно предположить, что $\big(X_{ij} | Y_i = t\big)$ имеет распределение Пуассона $\text{Pois}(\lambda)$ и оценить параметр λ методом максимального правдоподобия по подвыборке **из всех** X_{ij} (а не только по $X_{ji} = x_j$), для которых $Y_i = t$.
- Например, если купившие подписку пользователи $Y_i = 1$ в среднем открывали приложение X_{ij} по $\overline{X}_{ji} = 10$ раз в неделю, то методом максимального правдоподобия получаем $\hat{\lambda} = 10$, а значит предполагаем, что $\big(X_{ij} = x_j | Y_i = t\big) \sim \text{Pois}(0.1)$.
- Исходя из полученного результата вероятность того, что купивший подписку пользователь заходил в приложение 9 раз, составит:

$$P(X_{ij} = 9 | Y_i = 1) = e^{-10} \frac{10^9}{9!} \approx 0.125$$

- **Преимущество** – вся информация об условном распределении содержится в единственном параметре λ, оцениваемом по всем $\big(X_{ji} | Y_i = t\big)$, что позволяет избежать фрагментации данных и за счет этого снижает дисперсию оценок.
- **Недостаток** – неправильно подобранная форма распределения (предположили Пуассона, а на самом деле геометрическое) может привести к существенному смещению оценок.

Особенности применения наивного Байесовского классификатора

Признаки из непрерывных распределений

- Предположим, что признак X_{ij} был получен из непрерывного распределения (доход, вес и т.д.) с функцией плотности $f(z)$.
- Обозначим через $f_t(z)$ условную функцию плотности $(X_{ij}|Y_i = t)$ и будем использовать ее вместо вероятности $P(X_{ij} = z|Y_i = t)$.
- **Проблема** – в отличие от вероятностей функция плотности $f_t(z)$ не стандартизирована к шкале от 0 до 1, поэтому вклад функции плотности в условную вероятность $P(Y_i = t|X_{ij} = z)$ при различных t может сильно варьироваться.
- **Решение** – првести распределения $(X_{ij}|Y_i = t)$ для всех t к единой дисперсии, что сделает более сопостатавимыми между собой $f_t(z)$ при различных t.
- Обычно $(X_{ij}|Y_i = t)$ стандартизируют к нулевому математическое ожиданию и единичной дисперсии за счет того, что из каждого наблюдение вычитают выборочное среднее и делят эту разницу на выборочное стандартное отклонение. Обе выборочные характеристики считаются по подвыборке из X_{ij}, для которых $Y_i = t$.
- Поскольку данные стандартизированы к нулевому математическому ожиданию и единичной дисперсии, то $f_t(z)$ также подбирают из распределения с единичной дисперсией и нулевым математическим ожиданием, например, стандартного нормального $\mathrm{N}(0, 1)$.

Особенности применения наивного Байесовского классификатора

Подбор формы непрерывного распределения признака

- Для получения точных оценок $f_z(t)$ необходимо верно подобрать форму соответствующего распределения, в противном случае среднеквадратическая ошибка оценок плотностей может оказаться достаточно велика вследствие смещения.
- Можно попробовать рассмотреть различные распределения и подобрать оптимальное на основании информационного критерия, например, AIC или BIC.
- Функция плотности $f_z(t)$ может иметь дополнительные параметры. Например, при использовании стандартизированного к единичной дисперсии распределения Стьюдента необходимо оценить число степеней свободы. Это можно сделать, например, при помощи метода максимального правдоподобия.
- Функции плотности при различных t не обязательно должны быть одинаковыми. Например, доход среди тех, у кого случился дефолт, может иметь распределение Стьюдента, а среди тех, у кого не случился – нормальное распределение.
- В качестве альтернативы параметрическому оцениванию функции плотности можно прибегнуть к непараметрическим методам оценивания, например, воспользовавшись ядерным оцениванием или гистограммой.

Особенности применения наивного Байесовского классификатора

Превращение непрерывных переменных в дискретные

- В случае возникновения сложностей с подбором $f_t(z)$ непрерывный признак X_{ij} можно превратить в дискретный за счет его разбиения на интервалы (**binning**).
- Например, непрерывную переменную на доход можно превратить дискретную, разделив индивидов на тех, кто зарабатывает менее 50 тысяч рублей, от 50 до 100 тысяч рублей и более 100 тысяч рублей.
- Такой подход часто критикуется, поскольку из-за разбиения непрерывной переменной на набор дискретных мы теряем часть полезной информации.
- Аналогичный подход возможен и в случае с непрерывными **целевыми** переменными. Например, вместо непосредственно дохода индивида можно предсказывать, в какую группу дохода он попадет.

Особенности применения наивного Байесовского классификатора

Сглаживание

- В случаях, когда $Y_i \in \{0, .., k\}$ принимает достаточно много значений, может возникать фрагментация данных, из-за которой условные вероятности $P(X_{ij} = z|Y_i = t)$ при некоторых t оцениваются по достаточно малому числу наблюдений.
- Особенно проблематичным является случай, когда $\hat{P}(X_{ij} = z|Y_i = t) = 0$ при некотором j, поскольку тогда $\hat{P}(Y_i = t|X_i = z) = 0$ независимо от оценок других вероятностей $\hat{P}(X_{iw} = z|Y_i = t)$, где $w \neq j$.
- В таких случаях применяют технику **сглаживания**, искусственно уменьшая большие вероятности и увеличивая малые.
- Наиболее популярно **сглаживание Лапласа**:

$$\hat{P}(X_{ij} = z|Y_i = t) = \frac{l + \hat{P}(Y_i = t) \prod_{j=1}^{m} \hat{P}(X_{ij} = x_j|Y_i = t)}{l \times (k+1) + \sum_{q=0}^{k} \hat{P}(Y_i = q) \prod_{j=1}^{m} \hat{P}(X_{ij} = x_j|Y_i = q)}$$

Где $l > 0$ это параметр, отражающий силу сглаживания.

- Даже если иногда $\hat{P}(X_{ij} = x_j|Y_i = t) = 0$, итоговая условная вероятность будет отличаться от нуля.
- Чем больше l, тем сильнее все вероятности сглаживаются к равномерному распределению. Обычно l берут небольшим, не более 5, чтобы он оказывал существенное влияние лишь в случаях, когда некоторые вероятности крайне близки к 0.