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:
  - Признаки и целевая переменая.
  - Классификатор и условная вероятность.
  - Проклятье размерности и фрагментация данных.
  - Переобучение.
  - Точность прогноза.
  - Обучающая и тестовая выборки.
  - Кросс-валидация.
  - Графические модели и направленные ациклические графы.
  - Априорная вероятность и факторы.

Байесовский классификатор

Основная идея

- Рассмотрим бизнес, который заинтересован в нахождении клиентов, готовых совершить покупку.
- У бизнеса имеется информация о **признаках** (характеристиках) клиентов, на основании которой необходимо определить потенциальных покупателей: факт покупки - бинарная **целевая переменная** (1 - купил, 0 - не купил).
- **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 являться оценкой **условной вероятности** 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.

- Имеется n-мерный вектор значений **целевой переменной** (target) Y и матрица **признаков** (features) 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 $p, p_j \in (0, 1)$.
- Например, в качестве бинарной целевой переменной Y_i можно рассматривать факт совершения покупки, наступления дефолта или ухода клиента. В роли признаков X_i могут выступать различные характеристики клиентов, такие как возраст, пол и образование.
- Если i-й клиент совершил покупку $Y_i = 1$, имеет высшее образование $X_{1i} = 1$ и не состоит в браке $X_{i2} = 0$, то $X_i = (X_{i1}, X_{i2}) = (1, 0)$.
- Параметры p, p_1 и p_2 в данном примере будут отражать **безусловные вероятности** покупки, наличия образования и наличия брака соответственно.
- В силу закона больших чисел состоятельная оценка $\hat{p}$ может быть получена как:

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

Байесовский классификатор

Оценивание условных вероятностей

- На практике нас обычно интересуют не безусловные, а условные вероятности, поскольку они позволяют уточнить информацию о вероятности принятия целевой переменной Y_i того или иного значения за счет информации о признаках X_i.
- Рассмотрим условную вероятность:

$$p_x = P(Y_i = 1 | X_i = x) = \frac{P(Y_i = 1, X_i = x)}{P(X_i)}, \text{ где } x \in \{0,1\}^m$$

- Например, $P(Y_i = 1 | X_i = (1,0,1))$ может отражать условную вероятность покупки $Y_i = 1$ для имеющего высшего образование $X_{i1} = 1$ холостого $X_{i2} = 0$ клиента с детьми $X_{i3} = 1$.
- Обозначим через $n(x)$ число наблюдений таких, что $X_i = x$, а через $n(x, y)$ число наблюдений таких, что $X_i = x$ и $Y_i = 1$.
- Например, $n(x)$ может отражать число молодых женщин с высшим образованием, а $n(x, y)$ - количество покупой среди таких клиентов.
- Поскольку $n(x, y)/n$ – это состоятельная оценка $P(X_i = x, Y_i = 1)$ и $n(x)/n$ – это состоятельная оценка $P(X_i = x)$, то по теореме Слуцкого состоятельная оценка p_x может быть получена как:

$$\hat{p}_x = \frac{n(x,y)/n}{n(x)/n} = \frac{n(x,y)}{n(x)} = \frac{\text{Число наблюдений таких, что } X_i = x \text{ и } Y_i = 1}{\text{Число наблюдений таких, что } X_i = x} = (\text{Доля } Y_i = 1 \text{ среди } X_i = x)$$

- Например, условная вероятность совершения покупки молодой женщиной с высшим образованием может быть оценена как доля покупок среди молодых женщин с высшим образованием.

- 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 $I(5 \geq 3) = 1$.
- Оценку условной вероятности можно записать с помощью функции индикатора:

$$\hat{p}_x = n(x, y)/n(x) = \sum_{i=1}^{n} I(X_i = x, Y_i = 1) / \sum_{i=1}^{n} I(X_i = x)$$

- **Бинарный классификатор** прогнозирует значение бинарной целевой переменной $y \in \{0, 1\}$ в зависимости от значения вектора признаков $x \in \{0, 1\}^m$, например, ориентуясь на значение условной вероятности:

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

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

Байесовский классификатор

Пример

- 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_i – прекратил ли клиент пользоваться приложением в прошлом месяце (1 – да, 0 – нет).
X_{i1} – сократилась ли частота использования приложения в позапрошлом месяце (1 – да, 0 – нет).
X_{i2} – изменились ли основные функции, используемые пользователем в позапрошлом месяце (1 – да, 0 – нет).

Данные

i	Y_i	X_{i1}	X_{i2}
1	0	1	0
2	1	1	1
3	1	1	0
4	1	0	1
5	0	1	1
6	1	0	1
7	0	0	0
8	1	1	0

- Пусть $x = (1, 0)$, тогда $n(x) = 3$ и $n(x, y) = 2$, откуда $\hat{p}_x = \hat{P}(Y_i = 1 | X_i = x) = 2/3$.
- 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

- **Вопрос** – с какой вероятностью совершит покупка женщина средних лет с высшим образованием, **тремя детьми и ипотекой**.
- Допустим, что из 100 женщин средних лет с высшим образованием 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 достаточно точно 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 условную вероятность покупки (характеристика распределения).
- **Решение** – ввести допущения об условной вероятности, которые позволят оценивать ее с помощью большего числа наблюдений.

Наивный Байесовский классификатор

Определение условной независимости двух случайных величин

- Дискретные случайные величины X_1 и X_2 независимы, если при любых $x_1, x_2 \in R$ соблюдается:

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1) \times P(X_2 = x_2)$$

- Дискретные случайные величины X_1 и X_2 независимы при условии случайной величины Y, если при любых x_1, $x_2 \in R$ и любом $y \in R$ таком, что $P(Y = y) \neq 0$, выполняется:

$$P(X_1 = x_1, X_2 = x_2 | Y = y) = P(X_1 = x_1 | Y = y) \times P(X_2 = x_2 | Y = y)$$

Пример - сперва вы подбрасываете честную монетку, результат выпадения которой является бернуллиевской случайной величиной Y (1 – орел, 0 – решка). Если выпадает орел, то затем вы два раза подбрасываете обычную монетку, а если решка – монетку, которая падает орлом с вероятностью 0.9. Результаты первого и второго бросков являются бернуллиевскими случайными величинами X_1 и X_2 (1 – орел, 0 – решка). Очевидно, что если известен результат предварительного броска Y, то результаты последующих бросков X_1 и X_2 условно независимы, откуда, например:

$$P(X_1 = 1, X_2 = 0 | Y = 1) = P(X_1 = 1 | Y = 1) \times P(X_2 = 0 | Y = 1) = 0.5 \times 0.5 = 0.25$$

$$P(X_1 = 1, X_2 = 0 | Y = 0) = P(X_1 = 1 | Y = 0) \times P(X_2 = 0 | Y = 0) = 0.9 \times 0.1 = 0.09$$

$$P(X_1 = 1, X_2 = 0) = P(X_1 = 1, X_2 = 0 | Y = 1)P(Y = 1) + P(X_1 = 1, X_2 = 0 | Y = 0)P(Y = 0) =$$

$$= 0.25 \times 0.5 + 0.09 \times 0.5 = 0.17 \textbf{ (по формуле полной вероятности)}$$

$$P(X_1 = 1)P(X_2 = 0) = (0.5 \times 0.5 + 0.9 \times 0.5) \times (0.5 \times 0.5 + 0.1 \times 0.5) = 0.21$$

$$P(X_1 = 1, X_2 = 0) \neq P(X_1 = 1)P(X_2 = 0) \implies X_1 \text{ и } X_2 \text{ зависимы}$$

Наивный Байесовский классификатор

Идея условной независимости

- Мы пытаемся спрогнозировать, 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:

$$P(\text{Купил игру}|\text{Женщина, Дорогая видеокарта}) = \quad \textbf{формула Байеса}$$

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

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

- 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.

Наивный Байесовский классификатор

Идея оценивания вероятностей

- Для краткости обозначим события: К - купил игру, Д - Дорогая видеокарта, Ж - женщина.
- Необходимо избавиться от проклятья размерности не только в числителе, но и в знаменателе условной вероятности, для чего применим формулу полной вероятности и воспользуемся допущением об условной независимости:

$$P(\text{Женщина}, \text{Дорогая видеокарта}) = P(\text{Ж}, \text{Д}) =$$

$$= \underbrace{P(\text{Ж}, \text{Д}|\text{К})P(\text{К}) + P(\text{Ж}, \text{Д}|\overline{\text{К}})P(\overline{\text{К}})}_{\text{формула полной вероятности}} = \underbrace{P(\text{Ж}|\text{К})P(\text{Д}|\text{К})P(\text{К}) + P(\text{Ж}|\overline{\text{К}})P(\text{Д}|\overline{\text{К}})P(\overline{\text{К}})}_{\text{допущение об условной независимости}}$$

- Мы вновь расписали совместную вероятность через маржинальные условные вероятности, оценивание которых возможно, как правило, по большому числу наблюдений.
- В результате оценка условной вероятности принимает вид:

$$\hat{p}_x = \hat{P}(\text{Купил игру}|\text{Женщина, Дорогая видеокарта}) = \frac{\hat{P}(\text{Ж}|\text{К})\hat{P}(\text{Д}|\text{К})\hat{P}(\text{К})}{\hat{P}(\text{Ж}|\text{К})\hat{P}(\text{Д}|\text{К})\hat{P}(\text{К}) + \hat{P}(\text{Ж}|\overline{\text{К}})\hat{P}(\text{Д}|\overline{\text{К}})P(\overline{\text{К}})}$$

- Полученная условная вероятность ложится в основу **наивного байесовского классификатора**, который отличается от байесовского классификатора лишь способом расчета условных вероятностей.
- Прогнозирование осуществляется по аналогии $\hat{y}(x) = I(\hat{p}_x > 0.5)$.

Наивный Байесовский классификатор

Пример 1

- Представим данные о покупках игры в форме таблиц:

Купил игру

	Мужчина	Женщина
Дорогая видеокарта	25	6
Не дорогая видеокарта	50	14

Не купил игру

	Мужчина	Женщина
Дорогая видеокарта	3	10
Не дорогая видеокарта	22	70

- Байесовский классификатор оценивает условную вероятность покупки игры женщиной с дорогой видеокартой как $\hat{p}_x = \hat{P}(\text{К}|\text{Ж}, \text{Д}) = 6/(10+6) = 0.375$.
- При допущении об условной независимости эта вероятность может быть оценена иначе:

$$\hat{P}(\text{К}) = \frac{25+50+6+14}{(25+50+6+14)+(3+22+10+70)} = \frac{19}{40} \implies \hat{P}(\overline{\text{К}}) = \frac{21}{40}$$

$$\hat{P}(\text{Ж}|\text{К}) = \frac{6+14}{25+50+6+14} = \frac{4}{19} \qquad \hat{P}(\text{Ж}|\overline{\text{К}}) = \frac{10+70}{3+22+10+70} = \frac{16}{21}$$

$$\hat{P}(\text{Д}|\text{К}) = \frac{25+6}{25+50+6+14} = \frac{31}{95} \qquad \hat{P}(\text{Д}|\overline{\text{К}}) = \frac{3+10}{3+22+10+70} = \frac{13}{105}$$

$$\hat{P}(\text{К})\hat{P}(\text{Ж}|\text{К})\hat{P}(\text{Д}|\text{К}) = \frac{19}{40}\frac{4}{19}\frac{31}{95} = \frac{31}{950} \qquad \hat{P}(\overline{\text{К}})\hat{P}(\text{Ж}|\overline{\text{К}})\hat{P}(\text{Д}|\overline{\text{К}}) = \frac{21}{40}\frac{16}{21}\frac{13}{105} = \frac{26}{525}$$

$$\hat{p}_x = (31/950)/(31/950 + 26/525) = 651/1639 \approx 0.397$$

Наивный Байесовский классификатор

Условная независимость случайных величин

- Носителем m-мерного случайного вектора X с компонентами из дискретных распределений называется множество значений, которое этот вектор принимает с ненулевой вероятностью:

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

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

$$P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m|Y = y) = \prod_{i=1}^{m} P(X_i = x_i|Y = y)$$

Наивный Байесовский классификатор

Оценивание вероятностей при допущении об условной независимости

- 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 $Y_i = t$.
- **Преимущество** – 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.
- **Наивный Байесовский классификатор** отличается от байесовского тем, что оценивает условные вероятности при допущении об условной независимости.

Наивный Байесовский классификатор

Альтернативная запись классификатора

- **Проблема 1** – для того, чтобы рассчитать условную вероятность p_x, необходимо посчитать как числитель, так и знаменатель, что может быть неэффективно с точки зрения временных затрат.
- **Проблема 2** – при расчете произведения большого числа вероятностей на компьютере часто возникают существенные численные погрешности (поскольку это произведение большого числа малых чисел), для уменьшения которых произведение вероятностей желательно заменить на сумму логарифмов.
- **Решение** – обратим внимание, что $p_x = P(Y_i = 1|X_i)$ и $(1 - p_x) = P(Y_i = 0|X_i)$ имеют одинаковый знаменатель. Следовательно, классифицирующее правило можно сформулировать как:

$$\hat{y}(x) = I(\hat{p}_x > 1 - \hat{p}_x) = I(\hat{p}_x / (1 - \hat{p}_x) > 1) = I\left(\frac{\hat{P}(Y_i = 1)\prod_{j=1}^{m}\hat{P}(X_{ij} = x_j|Y_i = 1)}{\hat{P}(Y_i = 0)\prod_{j=1}^{m}\hat{P}(X_{ij} = x_j|Y_i = 0)} > 1\right) =$$

$$= I\left(\ln \hat{P}(Y_i = 1) + \sum_{j=1}^{m} \ln \hat{P}(X_{ij} = x_j|Y_i = 1) > \ln \hat{P}(Y_i = 0) + \sum_{j=1}^{m} \ln \hat{P}(X_{ij} = x_j|Y_i = 0)\right)$$

- **Примечание** – данная формула не работает, если хотя бы одна из вероятностей равняется 0, поскольку $\ln(0)$ не определен. Однако, на практике при большом числе наблюдений у наивного Байесовского классификатора такие случаи встречаются крайне редко.

- Рассмотрим задачу прогнозирования 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:

$$\hat{y}(x) = I\left(\frac{\hat{P}(Y_i = 1)\hat{P}(X_{i1} = 1|Y_i = 1)\hat{P}(X_{i2} = 0|Y_i = 1)}{\hat{P}(Y_i = 0)\hat{P}(X_{i1} = 1|Y_i = 0)\hat{P}(X_{i2} = 0|Y_i = 0)} \geq 1\right) = I\left(\frac{\frac{3}{5} \times \frac{2}{3} \times \frac{2}{3}}{\frac{2}{5} \times \frac{1}{2} \times \frac{1}{2}} \geq 1\right) = I\left(\frac{8}{3} \geq 1\right) = 1$$

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

Оценивание качества моделей

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 **обучается** модель, то есть оцениваются ее параметры (например, факторы и априорные вероятности).
- **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.

- Применим Байесовский классификатор на следующих данных:

X_{i1}	1	1	0	0	0	1
X_{i2}	1	0	1	0	1	0
Y_i	1	1	0	0	0	0
$\hat{Y}_i$	1	1	0	0	0	1
Выборка	Обучающая				Тестовая	

- Точности на обучающей $\text{ACC}_{\text{train}} = 1$ и тестовой $\text{ACC}_{\text{test}} = 0.5$ выборках.
- Если в обучающей выборке не более чем по 1 разу встречаются все возможные комбинации признаков, то Байесовский классификатор будет давать идеальную точность на этой выборке.
- Однако, в таком случае мы оцениваем каждую условную вероятность по очень малому числу наблюдений, а значит ее оценка будет ненадежна, поскольку обладает большой дисперсией.
- **Вывод** – увеличивая число признаков, мы можем сделать Байесовский классификатор сколь угодно точным на обучающей, но не на тестовой выборке.
- Если точность прогнозов на обучающей выборке гораздо выше, чем на тестовой, то говорят, что модель **переобучилась**.

Оценивание качества моделей

Связь фрагументации данных и переобучения Байесовского классификатора

Рассмотрим, как по мере увеличения фрагментации данных будет изменяться точность прогнозов факта наличия работы, полученных с помощью Байесовского классификатора на подвыборках.

	Обучающая выборка			**Тестовая выборка**		
	Всего	Из них работают	ACC	Всего	Из них работают	ACC
Колчиество индивидов	2000	1100	0.55	1000	500	0.5
- из них мужчины	1000	600	0.6	550	308	0.56
- из них мужчины в браке	600	480	0.8	363	259	0.75
- из них мужчины, в браке, со степенью кандидата наук	50	45	0.9	20	16	0.8
- из них мужчины, в браке, со степенью кандидата наук по математике	2	2	1	1	0	0

Примечание – точности ACC указаны не для всей модели, а по подвыборкам с соответствующими значеинями признаков.

- **Проблема** – 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.
- **Решение** – 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:

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

- 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).

Пример с 5-частной кросс валидацией.

	Часть 1	Часть 2	Часть 3	Часть 4	Часть 5		
Раунд 1	Тест	Обучение	Обучение	Обучение	Обучение	$\rightarrow$	ACC_1
Раунд 2	Обучение	Тест	Обучение	Обучение	Обучение	$\rightarrow$	ACC_2
Раунд 3	Обучение	Обучение	Тест	Обучение	Обучение	$\rightarrow$	ACC_3
Раунд 4	Обучение	Обучение	Обучение	Тест	Обучение	$\rightarrow$	ACC_4
Раунд 5	Обучение	Обучение	Обучение	Обучение	Тест	$\rightarrow$	ACC_5

$$ACC_{CV} = \frac{ACC_1 + ACC_2 + ACC_3 + ACC_4 + ACC_5}{5}$$

Вместо ACC могут использовать и другие меры качества модели, такие как F1-score, AUC и зачение функции потерь, с которыми мы познакомимся позже.

- Рассмотрим пример с k-частной кросс валидации $k = 3$ навиного байесовского классификатора по $n = 6$ наблюдения. Наблюдения были случайным образом разделены на 3 части (folds).

Параметры	Части 2 и 3	Части 1 и 3	Части 1 и 2
$\hat{P}(Y_i = 1)$	1/4	1/2	3/4
$\hat{P}(X_{1i} = 1 \mid Y_i = 1)$	1	1/2	2/3
$\hat{P}(X_{2i} = 1 \mid Y_i = 1)$	0	1	2/3
$\hat{P}(X_{1i} = 1 \mid Y_i = 0)$	1/3	1/2	0
$\hat{P}(X_{2i} = 1 \mid Y_i = 0)$	1/3	1/2	0

i	Y	X_1	X_2	Часть	$\hat{p}_x$	$\hat{Y}$
1	1	1	1	1	0	0
2	1	0	1	1	0	0
3	1	1	0	2	0	0
4	0	0	0	2	0	0
5	0	1	1	3	1	1
6	0	0	0	3	1/4	0

- Чтобы спрогнозировать значение, например, для 3-го наблюдения, обратим внимание, что оно относится ко 2-му фолду, а значит при прогнозировании необходимо использовать оценки факторов, полученные по обучающей выборке из 1-го и 3-го фолдов, откуда:

$$\hat{p}_{(1,0)} = \frac{(1/2) \times (1/2) \times (1-1)}{(1/2) \times (1/2) \times (1-1) + (1-1/2) \times (1/2) \times (1-1/2)} = 0 \implies \hat{Y}_3 = I(\hat{p}_{(1,0)} > 0.5) = 0$$

- Нетрудно рассчитать точности как по фолдам, так и итоговую:

$$\text{ACC}_1 = (I(\hat{Y}_1 = Y_1) + I(\hat{Y}_2 = Y_2))/2 = (0+0)/2 = 0 \qquad \text{ACC}_2 = (I(\hat{Y}_3 = Y_3) + I(\hat{Y}_4 = Y_4))/2 = (0+1)/2 = 0.5$$

$$\text{ACC}_3 = (I(\hat{Y}_5 = Y_5) + I(\hat{Y}_6 = Y_6))/2 = (0+1)/2 = 0.5 \qquad \text{ACC} = (\text{ACC}_1 + \text{ACC}_2 + \text{ACC}_3)/3 = (0+0.5+0.5)/3 = 1/3$$

- Допущение об условной независимости в наивном Байесовском классификаторе является весьма нереалистичным для большинства встречающихся на практике задач.
- Например, в задаче прогнозирования дефолта заемщика образование и доход будут зависимыми и для тех, у кого наступил дефолт, и для тех, у кого он не наступил.
- В примере с прогнозированием покупки компьютерной игры среди тех, кто не купил игру, также могут быть геймеры. В таком случае допущение о независимости пола и наличия дорогой видеокарты может быть поставлено под сомнение в случае, когда в качестве условия рассматривается отсутствие покупки игры.
- 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.

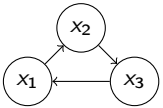


Рис.: Граф с циклом

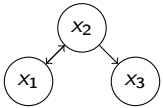


Рис.: Граф с двунаправленным ребром

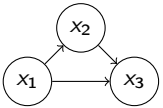


Рис.: DAG

- 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.
- **Примечание** – далее Y опущен для краткости и может быть любым из X. Также, для краткости опускается индекс наблюдений i, поэтому X_j отражает j-ю из случайных величин.

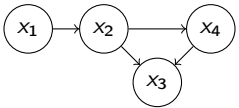


Рис.: DAG

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):

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

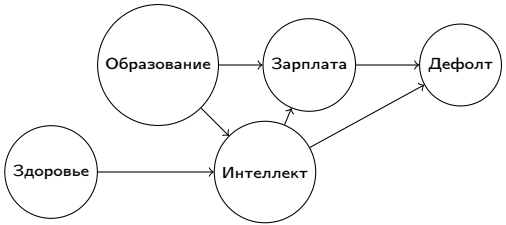
- Знаменатель условных вероятностей можно рассчитать с помощью формулы полной вероятности:

$$P(X_i|X_1,...,X_{i-1},X_{i+1},...,X_n) = \frac{P(X_1,..,X_n)}{\sum_{x\in \text{supp}(X_i)} P(X_1,...,X_{i-1},X_i=x,X_{i+1},...,X_n)} \propto$$

$$\propto P(X_i|\text{Parents}(X_i)) \prod_{j\neq i} P(X_j|\text{Parents}(X_j)) \propto P(X_i|\text{Parents}(X_i)) \prod_{Z\in \text{Children}(X_i)} P(Z|\text{Parents}(Z))$$

- Выражения, которым пропорциональна условная вероятность, упрощают формулу для классификатора и были получены за счет сокращения одинаковых знаменателей и условных вероятностей $P(X_j|\text{Parents}(X_j))$, не зависящих от значения X_i, то есть когда $X_i \notin \text{Parents}(X_j)$, что эквивалентно $X_j \notin \text{Children}(X_i)$.
- 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.

- Узел a называется **наследником** узла b, если из узла b, следуя по навлениям ребер (по стрелочкам), можно прийти в узел a. То есть наследниками узла b 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.
- 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. Наследниками образования являются зарплата, интеллект и дефолт, а не является – здоровье.

- 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

↓

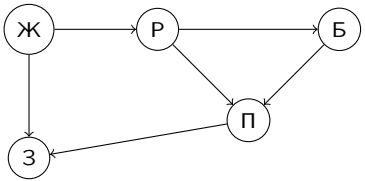
В случае необходимости улушается структура DAG, например, с помощью hill climb search

↓

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

Байесовские сети

Пример расчета совместных и условных вероятностей



Покупка (П)	1	0	1	0	1	1
Женщина (Ж)	1	1	1	1	0	1
Работа (Р)	1	0	1	1	0	1
Брак (Б)	1	0	1	1	1	1
Здоровье (З)	1	0	1	1	0	1

- Оценим совместные вероятности (для краткости А обозначает $A = 1$, а $\overline{A}$ обозначает $A = 0$):

$$\hat{P}(Ж, Р, Б, З, П) = \underbrace{\hat{P}(Ж)\hat{P}(Р|Ж)\hat{P}(Б|Р)}_{\text{сокращается}}\hat{P}(З|Ж, П)\hat{P}(П|Р, Б) = \underbrace{(5/6) \times 0.8 \times 1}_{\text{сокращается}} \times \underbrace{1 \times 0.75}_{0.75} = 0.5$$

$$\hat{P}\left(Ж, Р, Б, З, \overline{П}\right) = \underbrace{\hat{P}(Ж)\hat{P}(Р|Ж)\hat{P}(Б|Р)}_{\text{сокращается}}\hat{P}(З|Ж, \overline{П})\hat{P}(\overline{П}|Р, Б) = \underbrace{(5/6) \times 0.8 \times 1}_{\text{сокращается}} \times \underbrace{0.5 \times 0.25}_{0.125} = 1/12$$

- Оценим условную вероятность:

$$\hat{P}(П|Ж, Р, Б, З) = \frac{\hat{P}(П, Ж, Р, Б, З)}{\hat{P}(Ж, Р, Б, З)} = \frac{\hat{P}(П, Ж, Р, Б, З)}{\hat{P}(П, Ж, Р, Б, З) + \hat{P}(\overline{П}, Ж, Р, Б, З)} = \frac{0.5}{0.5 + 1/12} = \frac{0.75}{0.75 + 0.125} = \frac{6}{7}$$

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 $P(Y_i = t | X_i = x)$ оценивается как:

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

- Прогнозируется 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:

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

Особенности применения наивного Байесовского классификатора

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.

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 прогнозировать, в какую группу дохода он попадет.