

Фамилия:.....

Имя:.....

Группа:.....

Задача №1

Имеются следующие данные о заемщиках:

Дефолт	1	1	1	1	1	0	0	0	0	0
Образование	1	0	0	0	1	1	1	1	1	1
Брак	0	1	0	0	0	0	1	1	1	0

В **нейтральный к риску** банк пришел образованный **клиент**, состоящий в браке. Если банк не выдаст клиенту кредит, то ничего не получит, но и ничего не потеряет. Если банк выдаст клиенту кредит и клиент сможет выплатить этот кредит, то банк получит 100 рублей. Если банк выдаст клиенту кредит, но клиент не сможет вернуть этот кредит (произойдет дефолт), то банк потеряет 500 рублей.

1. С помощью **наивного** Байесовского классификатора оцените условную вероятность дефолта клиента (**5 баллов**)
2. Банк выдает кредит тогда и только тогда, когда наивный Байесовский классификатор прогнозирует отсутствие дефолта. Укажите цены различных видов прогнозов: TP – истинный положительный, FP – ложный положительный, TN – истинный отрицательный, FN – ложный отрицательный. (**5 баллов**)
3. Ориентируясь на информацию о ценах прогнозов, а также на оцененную условную вероятность дефолта, сделайте вывод о целесообразности выдачи кредита клиенту. (**5 баллов**)
4. Дефолт прогнозируется в случае, когда оценка его условной вероятности превышает 0.5. Вы использовали некоторый классификатор, который дал вам следующие оценки условных вероятностей:

$$\hat{P}(\text{Дефолт} = 1 | \text{Образование} = 1, \text{Брак} = 1) = 0.2$$

$$\hat{P}(\text{Дефолт} = 1 | \text{Образование} = 1, \text{Брак} = 0) = 0.4$$

$$\hat{P}(\text{Дефолт} = 1 | \text{Образование} = 0, \text{Брак} = 1) = 0.6$$

$$\hat{P}(\text{Дефолт} = 1 | \text{Образование} = 0, \text{Брак} = 0) = 0.8$$

Запишите (в форме таблицы) прогнозы дефолта для каждого заемщика в данных и укажите, к какому виду относятся эти прогнозы (TP, TN, FP или FN). Составьте матрицу путаницы (confusion matrix) и рассчитайте F1-метрику (F1-score) данного классификатора на данных. (**5 баллов**)

5. Изобразите наивный Байесовский классификатор как частный случай Байесовской сети: используйте ориентированный ациклический граф (DAG), чтобы описать связи между дефолтом, образованием и браком. (**10 баллов**)
6. Вы собираетесь обучить Байесовскую сеть и предполагаете, что образование является причиной брака, а брак – причиной дефолта. С помощью ориентированного ациклического графа (DAG) изобразите структуру данной Байесовской сети и, учитывая накладываемые ею ограничения на совместное распределение целевой переменной и признаков, оцените условную вероятность дефолта клиента. Объясните, какие факторы достаточно знать для расчета этой условной вероятности и сделайте вывод о целесообразности учета всех признаков при оценивании условных вероятностей дефолта. (**10 баллов**)

Решение:

1. Оценим априорные вероятности и факторы:

$$\begin{aligned}\hat{P}(\text{Дефолт} = 1) &= 0.5 & \hat{P}(\text{Дефолт} = 0) &= 0.5 \\ \hat{P}(\text{Образование} = 1 | \text{Дефолт} = 1) &= 0.4 & \hat{P}(\text{Образование} = 1 | \text{Дефолт} = 0) &= 1 \\ \hat{P}(\text{Брак} = 1 | \text{Дефолт} = 1) &= 0.2 & \hat{P}(\text{Брак} = 1 | \text{Дефолт} = 0) &= 0.6\end{aligned}$$

При допущении об условной независимости, то есть о независимости между образованием и браком при фиксированном дефолте, получаем:

$$\begin{aligned}\hat{P}(\text{Дефолт} = 1, \text{Образование} = 1, \text{Брак} = 1) &= \\ = \hat{P}(\text{Дефолт} = 1) \hat{P}(\text{Образование} = 1, \text{Брак} = 1 | \text{Дефолт} = 1) &= \\ = \hat{P}(\text{Дефолт} = 1) \hat{P}(\text{Образование} = 1 | \text{Дефолт} = 1) \hat{P}(\text{Брак} = 1 | \text{Дефолт} = 1) &= \\ = 0.5 \times 0.4 \times 0.2 &= 0.04\end{aligned}$$

По аналогии имеем:

$$\begin{aligned}\hat{P}(\text{Дефолт} = 0, \text{Образование} = 1, \text{Брак} = 1) &= \\ = \hat{P}(\text{Дефолт} = 0) \hat{P}(\text{Образование} = 1, \text{Брак} = 1 | \text{Дефолт} = 0) &= \\ \hat{P}(\text{Дефолт} = 0) \hat{P}(\text{Образование} = 1 | \text{Дефолт} = 0) \hat{P}(\text{Брак} = 1 | \text{Дефолт} = 0) &= \\ = 0.5 \times 1 \times 0.6 &= 0.3\end{aligned}$$

При допущении об условной независимости оценим условную вероятность дефолта:

$$\hat{P}(\text{Дефолт} = 1 | \text{Образование} = 1, \text{Брак} = 1) = \frac{0.04}{0.04 + 0.3} = \frac{2}{17} \approx 0.118$$

2. Запишем различные виды прогнозов и соответствующие им цены:

- а) TP - спрогнозировали дефолт и он бы наступил, $P_{TP} = 0$.
- б) TN - спрогнозировали отсутствие дефолта и он бы не наступил, $P_{TN} = 100$.
- в) FP - спрогнозировали дефолт, но он бы не наступил, $P_{FP} = 0$.
- г) FN - спрогнозировали отсутствие дефолта, но он бы наступил, $P_{FN} = -500$.

3. Ожидаемый выигрыш банка в случае с рассматриваемым клиентом составит:

$$\text{profit} = 100 \times \left(1 - \frac{2}{17}\right) - 500 \times \frac{2}{17} \approx 29.41$$

Поскольку ожидаемая полезность банка от выдачи кредита превышает ожидаемую полезность от отсутствия выдачи, то целесообразно выдать кредит.

4. Запишем таблицу с прогнозами и их видами:

Дефолт	1	1	1	1	1	0	0	0	0	0
Образование	1	0	0	0	1	1	1	1	1	1
Брак	0	1	0	0	0	0	1	1	1	0
Дефолт	0	1	1	1	0	0	0	0	0	0
Вид	FN	TP	TP	TP	FN	TN	TN	TN	TN	TN

Составим матрицу путаницы:

	$\widehat{\text{Дефолт}} = 1$	$\widehat{\text{Дефолт}} = 0$
Дефолт = 1	TP = 3	FN = 2
Дефолт = 0	FP = 0	TN = 5

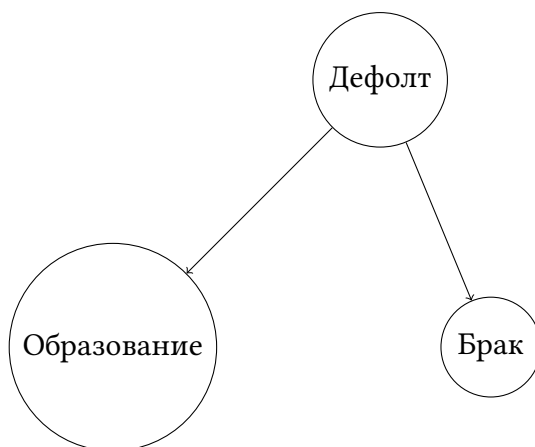
Посчитаем точность и полноту:

$$\text{precision} = \frac{3}{3+0} = 1 \quad \text{recall} = \frac{3}{3+2} = 0.6$$

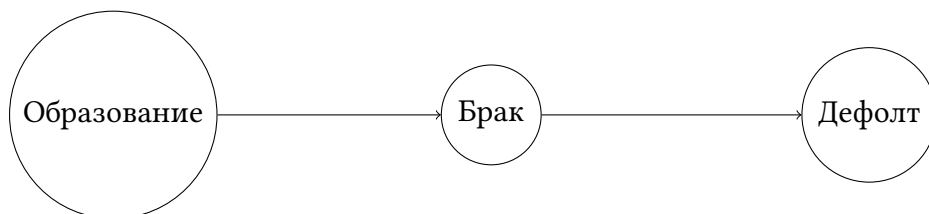
В результате получаем значение метрики:

$$F1 = 2 \times \frac{1 \times 0.6}{1 + 0.6} = \frac{6}{8} = 0.75$$

5. В наивном Байесовском классификаторе в качестве факторов рассматриваются $P(\text{Образование}|\text{Дефолт})$ и $P(\text{Брак}|\text{Дефолт})$. Следовательно, дефолт является родителем образования и брака, а значит наивный Байесовский классификатор можно изобразить следующим образом:



6. Иллюстрируем структуру Байесовской сети:



Сперва оценим совместные вероятности:

$$\begin{aligned}
 & \hat{P}(\text{Дефолт} = 1, \text{Образование} = 1, \text{Брак} = 1) = \\
 & \hat{P}(\text{Образование} = 1) \hat{P}(\text{Брак} = 1 | \text{Образование} = 1) \hat{P}(\text{Дефолт} = 1 | \text{Брак} = 1) = \\
 & = \underbrace{\frac{7}{10} \times \frac{3}{7}}_{\text{сокращается}} \times \frac{1}{5} = \frac{3}{50} = 0.075 \\
 & \hat{P}(\text{Дефолт} = 0, \text{Образование} = 1, \text{Брак} = 1) = \\
 & \hat{P}(\text{Образование} = 1) \hat{P}(\text{Брак} = 1 | \text{Образование} = 1) \hat{P}(\text{Дефолт} = 0 | \text{Брак} = 1) = \\
 & = \underbrace{\frac{7}{10} \times \frac{3}{7}}_{\text{сокращается}} \times \frac{3}{5} = \frac{9}{50} = 0.225
 \end{aligned}$$

Очевидно, что для расчета условной вероятности достаточно знать лишь оценку фактора $P(\text{Дефолт} = 0 | \text{Брак} = 1)$, поскольку остальные факторы сокращаются:

$$\hat{P}(\text{Дефолт} = 1 | \text{Образование} = 1, \text{Брак} = 1) = \frac{\frac{3}{50}}{\frac{3}{50} + \frac{9}{50}} = \frac{\frac{1}{5}}{\frac{1}{5} + \frac{3}{5}} = 0.25$$

Таким образом, для оценивания условной на брак и образование вероятности дефолта при соответствующей структуре Байесовской сети достаточно учитывать лишь брак.

Фамилия:.....

Имя:.....

Группа:.....

Задача №2

Исследователь оценивает эффект воздействия факта наличия высшего образования T_i на зарплату Y_i , в качестве контрольных переменных X_i рассматривая стаж и здоровье, а в качестве инструментальной переменной Z_i – образование родителей. Средний эффект воздействия оценивается следующим образом:

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{\hat{P}(T_i = 1 | X_i)} - \hat{E}(Y_i | X_i, T_i = 0)$$

1. Объясните, как можно использовать машинное обучение для расчета соответствующей оценки. Уточните, для чего в данном случае лучше использовать методы классификации, а для чего – регрессию. **(10 баллов)**
2. Сформулируйте допущение об условной независимости и выскажите содержательные (социально-экономические) соображения о возможных причинах нарушения данного допущения в рассматриваемом случае. Уточните, как нарушение этого допущения скажется на свойствах оценки \widehat{ATE} . **(10 баллов)**
3. Объясните, можно ли избежать негативных последствий нарушения допущения об условной независимости за счет тюнинга (подбора оптимальных гиперпараметров) методов машинного обучения. **(10 баллов)**
4. Назовите метод, не требующий сильных предположений о форме связи между контрольными переменными и целевой переменной, который можно применить для получения состоятельной оценки среднего локального эффекта воздействия при нарушении допущения об условной независимости. Укажите с помощью чего и какие именно два основных вида смещения позволяет ослабить соответствующий метод. **(15 баллов)**
5. При допущении об условной независимости исследователь также оценил несколькими методами условные эффекты воздействия $SATE_i$ и хочет выбрать лучший из методов, ориентируясь на следующую метрику качества оценок $SATE_i$:

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \hat{E}(Y_i | X_i, T_i) \right)^2$$

Объясните, можно ли считать данную метрику приемлимой. **(15 баллов)**

Решение:

1. Условную вероятность $P(T_i = 1|X_i)$ и условное математическое ожидание $E(Y_i = 1|X_i)$ можно оценить с помощью методов машинного обучения. При этом, для оценивания условных вероятностей уместно использовать классификацию (например, рашющие деревья или логистическую регрессию), а для условного математического ожидания – регрессию (например, регрессионные деревья).
2. Согласно допущению об условной независимости $E(Y_{ji}|X_i, T_i = j) = E(Y_{ji}|X_i)$, где $j \in \{0, 1\}$ и Y_{ji} отражает значение j -го потенциального исхода: Y_{1i} – зарплата людей при наличии высшего образования, Y_{0i} – зарплата людей при отсутствии высшего образования.

Способности индивида, которые он развил еще до поступления в университет, могут положительно влиять как на вероятность получения высшего образования, так и на заработную плату. Поскольку способности не входят в число контрольных переменных и могут существенно различаться между индивидами даже при одинаковом здоровье и стаже, это может приводить к нарушению допущения о независимости, так как, вероятно, $E(Y_{ji}|X_i, T_i = 1) > E(Y_{ji}|X_i)$, поскольку среди людей с высшим образованием могут чаще встречаться развитые способности. В результате оценка \widehat{ATE} окажется несостоятельной.

3. При нарушении допущения об условной независимости даже если вместо оценок $\hat{P}(T_i = 1|X_i)$ и $\hat{E}(Y_i|X_i, T_i)$ используются истинные значения $P(T_i = 1|X_i)$ и $E(Y_i|X_i, T_i)$, мы получаем:

$$\widehat{ATE} \xrightarrow{P} E(Y_{1i}|T_i = 1) - E(Y_{0i}|T_i = 0) \neq ATE$$

Поскольку качество используемых методов машинного обучения может повлиять лишь на точность оценок $\hat{P}(T_i = 1|X_i)$ и $\hat{E}(Y_i|X_i, T_i)$, то оценка \widehat{ATE} останется несостоятельной независимо от того, насколько хорошо будут подобраны гиперпараметры методов, используемых для оценивания условных вероятностей и условного математического ожидания.

4. В качестве альтернативы исследователь может использовать двойное машинное обучение с инструментальными переменными. В данном методе ортогональность по Нейману и кросс-фиттинг позволяют ослабить смещения, вызванные регуляризацией и переобучением соответственно.
5. Данная метрика не является хорошим вариантом, поскольку модель, точнее оценивающая Y_i , не обязательно будет точнее оценивать $CATE_i$. Например, даже если оценки $\hat{E}(Y_i|X_i, T_i = 1)$ и $\hat{E}(Y_i|X_i, T_i = 0)$ обладают очень большим, но примерно одинаковым смещением, то их разница, то есть \widehat{CATE}_i , может оказаться весьма точной оценкой, поскольку смещения сократятся. В качестве альтернативы MSE можно использовать, например, метрику, опирающуюся на псевдо-исходы:

$$MSE^* = \frac{1}{n} \sum_{i=1}^n \left(\hat{Y}_i^* - \widehat{CATE}_i \right)^2$$

где:

$$\hat{Y}_i^* = Y_i \left(\frac{T_i}{\hat{P}(T_i = 1|X_i)} + \frac{1 - T_i}{1 - \hat{P}(T_i = 1|X_i)} \right)$$