Тестовое задание

Дмитрий Донецков (ddonetskov@gmail.com)

10 июля 2019 г.

Содержание

1. Теоретическая задача №1

Вопрос:

> Можно ли использовать ковариацию двух признаков для анализа зависимостей между ними?

Ковариация двух признаков или, более корректно, двух случайных переменных показывает насколько сильно они меняются *совместно*, она может принимать значения от $-\infty$ до $+\infty$ Если посмотреть на логику формулы вычисления эмпирического значения ковариации двух случайных переменных X и Y, то можно увидеть, что это сумма таких произведений, что каждое произведение получается тем больше, чем более "синхронно"два признака отклоняются от своих средних значений:

$$\text{Теоретическая ковариация: } cov(X,Y) = \mathbb{E}(X - \mathbb{E}\,X)(Y - \mathbb{E}\,Y) \tag{1}$$

$$\text{Выборочная ковариация: } cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

Таким образом, более высокие значения ковариации получаются при более "синхронных"колебаниях переменных, что несёт информацию об их связи, об их зависимости между собой. Но! данную информацию тяжело интерпретировать, особенно тяжело её использовать для сравнения силы зависимости между парами разных признаков, например, между двумя парами признаков: рост и вес человека, размер обуви и вес человека. При сравнении этих двух признаков на основе значений ковариаций невозможно ответить на вопрос, что сильнее связано с весом человека. Данное затруднение исходит из того, что в ковариации содержится размерность исходных признаков, она не свободна от природы измерения самих признаков, которая может быть разной. Поэтому, для сравнения силы связи между признаками используют нормализованное значение ковариации - корреляцию, которая уже свободна от размерностей, и принимает своё значение на интервале от -1 до 1 вне зависимости от того, множество каких значений могут иметь сами признаки.

Дополнение: Ковариацию теоретически можно использовать для анализа зависимостей признаков, если не меняется размерность исследуемых признаков, но это крайне неудобно на практике, поэтому, как правило, для сравнений используют корреляцию.

Что можно добавить в материал:

- таблицу сравнения свойств ковариации и корреляции,
- пояснения, почему ковариация всё равно остаётся очень важным понятием (например, для ковариационных матриц).

2. Теоретическая задача №2

Вопрос:

> Ваш подчинённый обучил модель и получил качество классификации 99%. Какие бы вопросы про данные и про модель вы бы задали, чтобы убедиться, что модель полезная? Можно считать, что программных багов нет.

Качество классификации в 99% выглядит очень хорошим. Естественно, точная интерпретация такой оценки зависит от используемой метрики. Если из контекста задачи неясно какая метрика была использована, то следует задать вопрос об этом вплоть до просьбы показать формулу или фрагмент кода, при помощи которой/которого производилось вычисление метрики. Далее, для ухода от рассмотрения специфики различных метрик (prediction, recall, AUC и т.п.), предположим, что идёт речь о доле объектов с верно распознанными классами для них (accuracy).

Итак, качество классификации в 99% выглядит настолько хорошим, что стоит перепроверить, что полученная модель не является слишком специфичной, что потом её будет нецелесообразно использовать на практике. Другими словами, что нет случая переобучения, смещения оценки, при котором модель будет малоценной для новых объектов. Для такой проверки можно задать следующие вопросы:

1. Как распределены объекты по классам? Как удостоверились, что классификатор действительно работает (выдаёт разные значения)? Здесь, необходимо понять, что нет крайне выраженного дисбаланса, когда, например, подавляющее большинство объектов относится к одному классу, а сам классификатор всегда выдаёт значение только одного класса. Другими словами, стоит провести проверку насколько полученная модель лучше dumb-модели. Также, стоит отметить, что в этом случае, несмотря на высокое качество классификации, такое решение "в лоб"может иметь высокую стоимость для оставшегося 1% случаев, делая модель экономически неэффективной.

2. Каким образом была проведена оценка модели? Здесь, необходимо понять, что оба множества объектов, используемые для построения модели (тренировки) и её последующей оценки (тестирования), были взяты таким образом, что каждое множество достаточно хорошо представляет собой множество всех возможных объектов (генеральную совокупность). Можно подсказать, про техники train/test split, k-fold cross-validation.

Что можно добавить в материал:

- пояснения по тому, что метрика accuracy - неполна, есть разные метрики оценки качества классификатора: precision, recall, F_1 score и т.п.; некоторые из них специфичны, а другие - были сконструированы, как общие.

3. Практическая задача

Задание:

> Напишите функцию one_hot_encode(), которая принимает на вход объект pandas.DataFrame и название столбца, и выполняет one-hot-кодирование этого столбца. Гарантируется, что переданный столбец категориальный. Функция должна возвращать новую таблицу (pandas.DataFrame), в которой отсутствует старый столбец, а новые добавлены в конец. Использовать готовые функции для решения этой задачи (например, pandas.get_dummies()) не разрешается.

Код функции и её тесты оформлены в виде Jupyter-ноутбука и доступны для просмотра по ссылке https://nbviewer.jupyter.org/github/ddonetskov/misc/blob/master/yp_ds/one_hot_encoding.ipynb#.

❗ Примечание: реализованная функция не покрывает всех возможных практических случаев, а именно:

- если есть пропущенные данные в указанном столбце, то функция возвращает исключение `NotImplementedError`, т.к. логика обработки пропущенных значений может разнится в зависимости от задачи,
- кодирование выполняется для всех значений в столбце, никакое значение не удаляется, что может порождать проблему с коллинеарностью для соответствующих моделей (например, для линейной регрессии).

4. Урок

5. Упражнение (кейс)

6. Открытые вопросы