НЕТОЛОГИЯ
групп

Блок

FEATURE ENGINEERING

ЕГОР САЧКО

Lead Data Scientist

Сбербанк

egsachko@gmail.com

sachkoe

МАТЕРИАЛЫ ПО БЛОКУ

МАТЕРИАЛЫ ПО БЛОКУ

1. **“Learning scikit-learn: Machine Learning in Python”**
Raul Garreta,
Guillermo Moncecchi,
2013,
Packt

2. **“Hands-On Machine Learning with Scikit-Learn and Tensorflow: Concepts, Tools and Techniques to Build Intelligent Systems”**
Geron, A., 2017, O’Reilly Media

3. **“Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists”**
Zheng, A., Casari, A.,
2018,
O’Reilly Media

4. blog.kaggle.com/ - No Free Hunch

Занятие 1

ПРОБЛЕМЫ КАЧЕСТВА И РАЗМЕРНОСТИ ДАННЫХ

ЦЕЛИ ЗАНЯТИЯ

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ СМОЖЕТЕ

1. Осуществлять поиск **подмножества признаков**
2. Использовать **регуляризацию**
3. Уменьшать **пространство фич**
4. Оценивать **значимость переменных**
5. Использовать **sklearn для Feature selection**

ЧТО БУДЕМ ОБСУЖДАТЬ

ПЛАН ЗАНЯТИЯ

1. Обзор домашнего задания
2. Первичный анализ данных
3. Оценка значимости переменных
4. Сокращение размерности пространства данных
5. Регуляризация

Часть 1-2

Обзор домашнего задания Первичный анализ данных

Training & Test

A

Разбиваем на Training и Test сеты как можно раньше

B

Data snooping bias

C

Training set — выбор, тренировка и тюнинг моделей

D

Testing set — оценка финальной модели

Балансировка данных

Перекос данных

- 90 % данных — класс A, 10 % данных — класс B
- Модель всегда отвечает A — accuracy 90 %

Как бороться? Часть методов

- Oversampling and undersampling
- Синтетические данные
- Другие метрики *AUC, F1-score*
- Другие способы machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset

Масштабирование и нормализация

Масштабирование

- Standard $x' = \frac{x - \bar{x}}{\bar{o}}$
- Min-Max $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$

Нормализация

L1, L2, ...

Трансформация данных

Feature
SELECTION

Feature
ENGINEERING

Заключение

1

Подготовка данных ≈ тренировка моделей

2

Поиск аномалий и способы их решений — только training set

3

Полученные решения применяются к данным в обучающую модель

- Test set
- Новые данные

Практика

АНАЛИЗ БАНКОВСКИХ ТРАНЗАКЦИЙ

Часть 3

Оценка значимости переменных

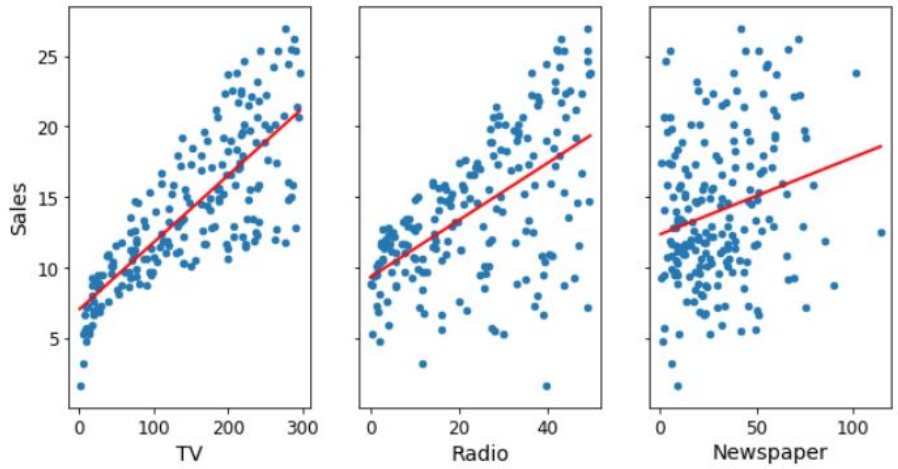
Датасет

- Продажи продукта ~ рекламные бюджеты на разные медиа
- Медиа: ТВ, радио и газеты
- URL: http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv

ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

Цель

Создать маркетинговый план на следующий год на основе данных из датасета, так, чтобы продажи продукта были высокими.



На какие вопросы пытаемся ответить?

- Есть ли связь между рекламным бюджетом и продажами?
- Насколько сильна связь между бюджетом и продажами?
  Можем ли мы предсказывать продажи на основе бюджета?
- Какие медиа способствуют продажам?
- Насколько точно мы можем предсказывать будущие продажи?
- Линейная ли зависимость между бюджетом и продажами?
- Есть ли эффект взаимодействия (synergy/interaction effect)
  между медийными бюджетами?

Линейная регрессия

$$sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper$$

- Предположим, что медийные бюджеты не зависят друг от друга
- Определим β_0, β_1, β_2, β_3

Standard error

- Интересно знать насколько точна наша аппроксимация

$$SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$$

- Доверительные интервалы
- Например, 95% значений β_0 будут в интервале [2.324,3.554]

Проверка гипотезы

- Нулевая и альтернативная гипотезы
  - H_0: между x_i и y **нет зависимости**
  - H_A: между x_i и y **есть зависимость**
- Для проверки гипотезы используется t-test

T-Statistics & P-value

$$t = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$$

- Если между x_i и y нет зависимости, то t соответсвует t-распределению с n−2 степенями свободы
- p-value - вероятность того, что при известном распределении наблюдаемое значение ≥ |t| (**при условии, что $\beta_i = 0$**)
- Если p-value достаточно маленький (< 1%), то мы можем отклонить H_0

Бюджеты и продажи

- 4 независимые гипотезы:
  - H_0: $\beta_i = 0$
  - H_A: $\beta_i \neq 0$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9389	0.312	9.422	0.000	2.324	3.554
TV	0.0458	0.001	32.809	0.000	0.043	0.049
Radio	0.1885	0.009	21.893	0.000	0.172	0.206
Newspaper	-0.0010	0.006	-0.177	0.860	-0.013	0.011

- Недостаток t-statistics: оценка важности каждого атрибута производится независимо от других

RSS & RSE

- RSS - Residual Sum of Squares $$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
- RSE - Residual Standard Error $$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$
  - p - количество predictor-ов
  - RSE штрафует модели, которым нужно больше predictor-ов для достижения одинаковых значения RSS

ОЦЕНКА ЗНАЧИМОСТИ ПЕРЕМЕННЫХ

R^2

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \quad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- TSS - Total Sum of Squares
- R^2 - показывает, какой процент вариативности (variance) объяснен моделью
- $R^2 \in [0, 1]$ - относительная величина, чем ближе к 1, тем лучше

F-Statistics

- Зависят ли продажи как минимум от одного из медиа ресурсов?
  - H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
  - H_A: как минимум один из $\beta_i \neq 0$
- Проверить такую гипотезу можно с помощью F-теста

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

F-Test

- Аналогичен t-тесту, только используется F-распределение
- F-тест для проверки равенства 0 только части параметров (q параметров из p)

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

RSS_0 - модель, в которой q параметров равны 0

Практика

АНАЛИЗ РЕКЛАМНЫХ БЮДЖЕТОВ

Часть 4

Сокращение размерности пространства данных

Сокращение размерности пространства данных

- Обучение с выбором подмножества фич (subset selection)
- Обучение с регуляризаций (shrinkage или regularization)
- Обучение с уменьшением размерности фич (dimensionality reduction)

Выбор подмножества фич

- Brute force (найти все комбинации, выбрать лучшую)
- Классические способы (эффективные)
  - Forward selection
  - Backward selection
  - Mixed selection

Forward stepwise selection

1. Рассмотрим модель M_0, которая не содержит предикторы
2. For k=0, ..., p-1:
    1. Рассмотрим p - k моделей, которые дополняют M_k одним дополнительным предиктором
    2. Выбираем лучшую модель среди p - k моделей (меньший RSS или больший R^2), назовем ее M_{k+1}
3. Выбираем лучшую модель из M_0, ..., M_p используя кросс-валидацию или метрики косвенной оценки

Backward stepwise selection

1. Рассмотрим модель M_p, которая содержит все предикторы
2. For k=p, p-1, ...,1:
    1. Рассмотрим k моделей, которые содержат k-1 предиктор модели M_k
    2. Выбираем лучшую модель среди k моделей (меньший RSS или больший R^2), назовем ее M_{k-1}
3. Выбираем лучшую модель из M_0, ..., M_p используя кросс-валидацию или метрики косвенной оценки

Mixed selection

- Работает как forward stepwise selection
- В конце каждом шага может сделать backward stepwise selection

Часть 5

Регуляризация

Регуляризация

- Линейная регрессия $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j * x_{ij})^2$
- Ridge $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j * x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$
- Lasso $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j * x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

ИТОГИ

1. Как оценивать **значимость переменных**
2. Как устроена **линейная регрессия**
3. Какие существуют типы **регуляризации**
4. Как осуществить **отбор признаков**

СПАСИБО ЗА ВНИМАНИЕ