Байесовское моделирование категориальных признаков

Надежда Бугакова

научный руководитель: к. ф.-м. н. И. Е. Кураленок

17 июня 2018 г.

СПб АУ НОЦНТ РАН

Задача машинного обучения с учителем

Существует:

$$(x, y) \sim P, x \in X, y \in Y$$

Классификация: $y_i \in \{0, 1\}$

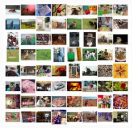


Рис. 1: Картинки

Хотим:

$$\phi : X \to y; \phi \in \mathbb{F}$$

$$\phi = \arg\min_{\phi \in \mathbb{F}} E_{(x,y)\sim P} L(\phi(x), y)$$

Рис. 2: Вещественные признаки

Обучающая выборка:

$$\{(x_i, y_i)\}_{i=1}^{m}, \text{реализации P.}$$

Ищем:

$$\hat{\phi} = \arg\min_{\phi \in \mathbb{F}} E_{\{(x_i,y_i)\}} L(\phi(x), y)$$

Рис. 3: Категориальные признаки

Категориальные признаки

- Нет отношения порядка.
- Количество различных значений бывает очень велико.

Методы:

- One-hot-encoding

- Специализированные под предметную область (например, использование вложенности признаков)
- Оценка параметров вероятностной модели (CatBoost)

Цели и задачи

Целью данной работы является изучение эффективности иерархического байесовского моделирования категориальных признаков для задачи классификации на основе обучающего множества.

Для достижения цели необходимо решить следующие **задачи**:

- Исследовать механизм работы с категориальными признаками, используемый в CatBoost.
- Разработать вероятностную модель для категориальных признаков с подбором параметров на основе обучающего множества.
- Реализовать преобразование категориальных признаков в упорядоченные.
- Сравнить на задачах обучения с учителем

Вероятностная модель для категориальных признаков

Идея: моделировать зависимость целевых значений от категориальных признаков.

$$y_c \sim Ber(\theta_c)$$

Простейщая модель: θ_c независимы.

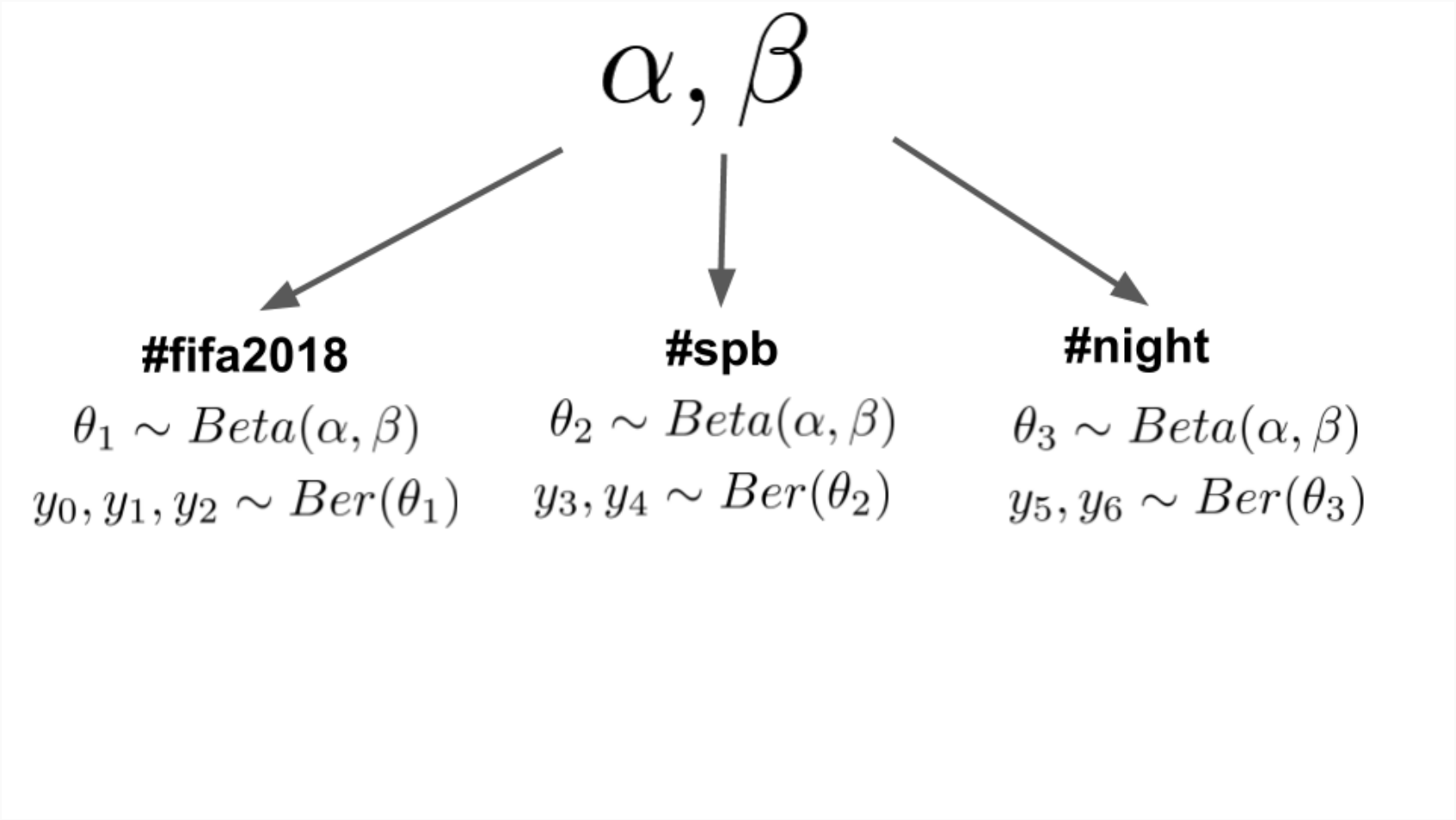
Проблема: переобучение.

Для категорий $y_0 \sim Ber(0.4); y_1 \sim Ber(0.5)$

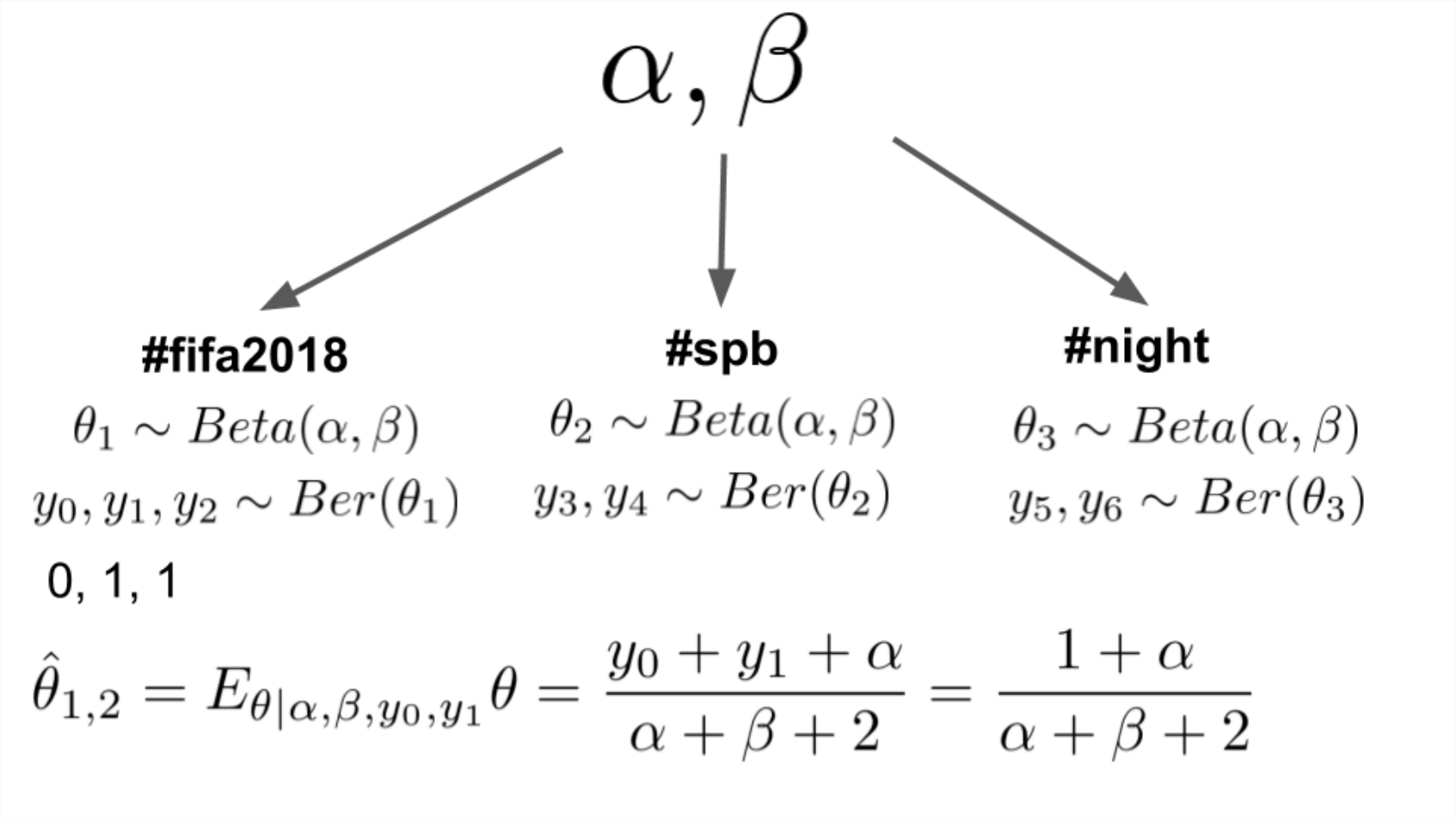
$$\text{Выборка: } \begin{array}{c|c} \text{cat0} & 0 \\ \hline \text{cat1} & 1 \end{array} \Rightarrow \begin{array}{c} \hat{\theta}_0 = \frac{0}{1} = 0 \\ \hline \hat{\theta}_1 = \frac{1}{1} = 1 \end{array}$$

Решение: моделирование со временем: оцениваем θ_c только на основе целевых значений ДО текущего момента.

Оценка параметров вероятностной модели в CatBoost



Оценка параметров вероятностной модели в CatBoost



Автоматический подбор

Максимизируем правдоподобие по (α, β):

$$\prod_c p(D_c|\alpha,\beta) = \prod_c \int_{\theta_c} \prod_{i=1}^{N_c} p(y_{c,i}|\theta_c) dBeta(\alpha,\beta)$$

$$\int_{\theta_c} \prod_{i=1}^{N_c} p(y_{c,i}|\theta_c) p(\theta_c|(\alpha,\beta)) d\theta_c = \int_{\theta_c} \frac{\theta_c^{\sum y_{c,i}+\alpha-1}(1-\theta_c)^{\sum(1-y_{c,i})+\beta-1}}{B(\alpha,\beta)} d\theta_c =$$

$$= \frac{B(\sum y_{c,i}+\alpha, \sum(1-y_{c,i})+\beta)}{B(\alpha,\beta)}$$

Реализация

Были реализованы на Python:

- Оптимизация априорных значений с помощью метода Ньютона на основе библиотеки scipy.
- Преобразование категориальных признаков в вещественные тремя способами:
  - С автоматическим подбором априорных значений;
  - С фиксированными в CatBoost априорными значениями;
  - С априорными значениями, где α рассчитывалась как среднее целевого значения по всей выборке и $\beta = 1$ - α.
- Воспроизводимые эксперименты.

Методика сравнения

- Gradient boosting (CatBoost).
- Бутстрепинг тестовой выборки + Wilcoxon signed-rank test.

Датасеты

Name	Atributes	Categorical	Learn size	Test size
adult	14	8	39074	9768
amazon	9	9	26215	6554
appet	419	38	40000	10000
kick	439	23	58388	14595

Тестирование

Уровень статистической значимости 0.01.

AUC (модель со временем):

Name	CatBoost+time	Auto+time	Simple+time
amazon	**0.8564**	0.853	0.855
adult	0.9275	**0.9281**	**0.9281**
appet	**0.8525**	0.8475	0.8468
kick	**0.7656**	**0.7657**	0.7638

AUC (модель без времени):

Name	CatBoost	Auto	Simple
amazon	0.8053	**0.8188**	0.81
adult	0.9292	0.9291	0.9292
appet	**0.6909**	0.6628	0.6615
kick	0.7561	**0.7705**	0.7623

Тестирование

Время обучения в секундах:

Name	CatBoost	Auto	Simple
amazon	48	38	35
adult	55	48	48
appet	261	226	227
kick	128	105	92

Name	CatBoost+time	Auto+time	Simple+time
amazon	46	38	37
adult	54	47	47
appet	267	235	236
kick	123	90	92

Результаты

- Исследован механизм преобразования категориальных признаков в CatBoost
- Придуман и изучен автоматический метод подбора априорных значений на основе обучающей выборки.
- Реализованы разные подходы к подбору априорных значений и произведено сравнение на наборе данных

Статьи и ссылки

GitHub (https://github.com/N-buga/SPBAU-CategBays)

- CatBoost (https://tech.yandex.com/catboost/)
- adult (https://archive.ics.uci.edu/ml/datasets/Adult)
- appet (http://www.kdd.org/kdd-cup/view/kdd-cup-2009/Data)
- amazon (https://www.kaggle.com/c/amazon-employee-access-challenge)
- kick (https://www.kaggle.com/c/DontGetKicked)