Прогнозирование направления движения цены биржевых инструментов по новостному потоку

Валентин Ахияров, Александр Борисов, Иван Говоров, Валентин Родионов

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов)/ ФУПМ ФРТК, весна 2019

Предсказание направления движения цены биржевых инструментов

Проблема

Проанализировать большой объём новостей на их влияние на цены.

Задача

Задача бинарной классификации направления движения разности двух временных рядов по отчётам.

Метод решения

Предлагается с помощью метода мешка слов представить новости в виде векторов. Затем с помощью неотрицательного матричного разложения получить матрицу для набора отчётов, на которой впоследствие обучать модели.

Литература

- **О рынке**
  Hongping Hu, Li Tang, Shuhua Zhang, Haiyan Wang (2018) *Predicting the direction of stock markets using optimized neural networks with Google Trends*, Neurocomputing.
- **О временных рядах**
  Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky (2014) *On the Importance of Text Analysis for Stock Price Prediction*, Proceedings of the Ninth International Conference on Language Resources and Evaluation.
  Anna Potapenko, Artem Popov, Konstantin Vorontsov (2017) *Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks*, CoRR.
- **О прогнозировании**
  Усманова К. Р., Кудияров С. П., Мартышкин Р. В., Замковои А. А., Стрижов В. В. (2018) *Анализ зависимостей между показателями при прогнозировании объема грузоперевозок*, Системы и средства информатики.

Постановка задачи

Дано

Отчеты 8K для 1500 компаний об их внутренних событиях с полями File, Time, Events, Text.

Задача

Бинарная классификация направления движения цен
$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, где $y_i \in \{0, 1\}$, 0 — stay, 1 — move.
Ответы $\mathbf{y}$ – бернуллиевский случайный вектор с плотностью
$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} p_i^{y_i}(1 - p_i)^{1-y_i}$

Функция ошибки:
$E(\mathbf{w}) = -\ln p(\mathbf{y}|\mathbf{w}) = -\sum_{i=1}^{m} y_i \ln p_i + (1 - y_i)\ln(1 - p_i)$

Требуется оценить вектор параметров $\hat{\mathbf{w}}$, доставляющий минимум функции ошибки: $\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^n} E(\mathbf{w})$

Решение

Алгоритм классификации имеет вид: $a(\mathbf{x}) = sign(\sigma(\mathbf{x}, \mathbf{w}) - \sigma_0)$ где σ_0 — задаваемое пороговое значение функции регрессии $p_i = \frac{1}{1+\exp(-\mathbf{x}_i^T \mathbf{w})} = \sigma(\mathbf{x}_i^T \mathbf{w}) \equiv \sigma_i$.

Отображение f является бинарной классификацией и отображает признаковое описание текста $\mathbf{x} \in \mathbf{x}(t)$ в метку класса $\{0, 1\}$:

$$f : (\mathbf{w}, \mathbf{x}) \mapsto y$$

Назовём вектор $\mathbf{w}$ вектором параметров классификатора. Требуется найти оптимальный классификатор $f(\mathbf{x}_i)$ при $\mathbf{x}_i \in \mathfrak{D}_t$ из условия:

$$\hat{f} = \text{argmin}_f S$$

где $S = \{S_1, S_2\}$.

Вычислительный эксперимент

Рассматриваемые модели

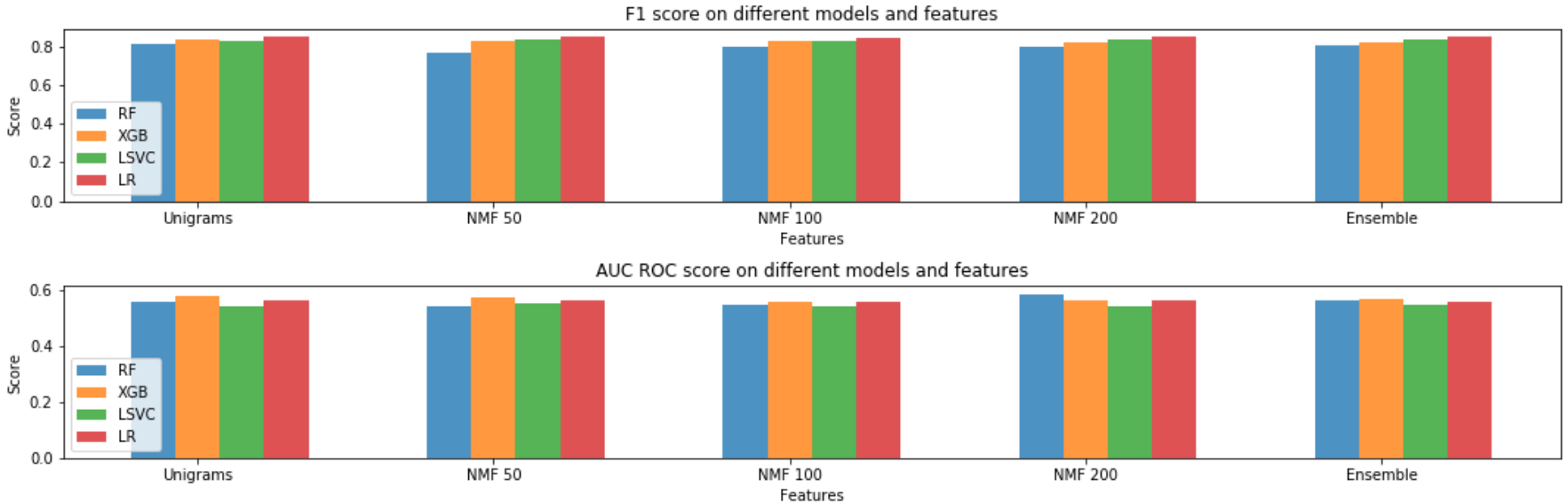
- Random Forest (RF)
- Logistical Regression (LR)
- Linear SVM (LSVC)
- XGBoost (XGB)

Критерии качества

- $F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$, где $Precision = \frac{TP}{TP+FP}$ — точность, а $Recall = \frac{TP}{TP+FN}$ — полнота.
- AUC-ROC (площадь под ROC-кривой)

Результаты

	Best Average			
Model	F1 Score	Features	AUC ROC	Features
RF	0.8093	Unigrams	**0.5838**	NMF 200
XGB	0.835	Unigrams	0.5805	Unigrams
LR	**0.8473**	NMF 200	0.5624	Unigrams
LSVC	0.8349	NMF 50	0.5495	NMF 50



Заключение

Результаты показали удовлетворительное решение поставленной задачи. Лучшей комбинацией была модель Random Forest на Unigrams со следующими гиперпараметрами:

- max_depth = None
- min_samples_leaf = 3
- min_samples_split = 5
- n_estimators = 2000

F1-score = 0.8615, AUC ROC = 0.5961 на трёх выборках (train, val, test)