Машинное обучение с учителем: базовые процедуры, сложности и возможности для социальных наук

Айгуль Мавлетова, НИУ ВШЭ

23 – 24 сентября 2016
Москва

Цель

Цель: дать общее представление о возможностях использования алгоритмов машинного обучения с учителем при работе с текстами.

План

1. Машинное обучение: обучение с учителем и обучение без учителя.
2. Text mining: базовые процедуры.
3. Построение моделей:

    - Обучающая выборка, тестовая выборка, перекрёстная проверка
    - Алгоритмы
    - Меры качества
    - Проблема переобучения
4. Эмпирический кейс

Машинное обучение – систематическое обучение алгоритмов и систем, в результате которого их знания или качество работы возрастают по мере накопления опыта.

Три основные составляющие:

- **Задачи:** чаще всего связано с классификацией объектов.
- **Модели:** результат алгоритмов, примененного к данным.
- **Признаки:** «язык», на котором описываются объекты предметной области.

Объекты представляются в виде признаков.

Supervised learning vs. unsupervised learning

Обучение с учителем **(supervised learning) – обучение по размеченным данным**

Обучение без учителя **(unsupervised learning) – обучение по неразмеченным данным**

Процедуры:

- Поиск и сбор информации.
- Предварительная обработка текстов.
- Применение алгоритмов.
- Интерпретация результатов.

Превращение текстовой информации в числовую

•Нам не важен порядок слов.

•Текст – это набор слов, которые встречаются в тексте с определенной частотой.

•**Униграммная модель**: только одно слово.

Предварительная обработка

Цель предварительной обработки текстов:
уменьшить количество уникальных слов в тексте, т.е. уменьшить размерность.

Какие стандартные процедуры используются?

1. Нормализация

Стемминг – преобразование слова до основы.

Экономически развитые регионы России сохранили в этом году наиболее высокие рейтинги качества жизни населения… Ивановской области вырасти в рейтинге помог уровень экономического развития, который, по данным исследования, увеличил прибыль компаний и снизил безработицу. Неизменным ростом по динамике прошлого года отличилась Рязанская область, которая поднялась с 35-го на 30-е место. Там позитивные изменения случились сразу по половине показателей, что пока нетипично для российских регионов, которые в большинстве, как отмечают эксперты, не могут удерживать лидирующие позиции повсеместно.

Количество уникальных слов уменьшилось с 71 до 65

2. Стоп-слова

3. Слова, которые редко встречаются.

4. Регистр: привести все слова к нижнему или ВЕРХНЕМУ регистру

Матрица терминов-документов **(document-term matrix)**

1. Подсчитать частоту встречаемости определенных слов.

Почти 70% выпускников лицея НИУ ВШЭ этого года стали студентами топ-10 вузов России, при этом практически каждый второй лицеист выбрал Вышку. Среди них – победители и призеры Всероссийской олимпиады школьников и олимпиады «Высшая проба».
Студентами Вышки стали 203 из 381 выпускников лицея 2016 года, или 53,2%. Самыми популярными факультетами среди них оказались факультет коммуникаций, медиа и дизайна (там учатся 47 выпускников лицея, или 22,7% от числа поступивших в Вышку), факультет социальных наук (29 выпускников, или 14,3%), факультет бизнеса и менеджмента (27 выпускников, или 13,3%), факультет мировой экономики и мировой политики (26 выпускников, или 12,8%) и факультет экономических наук (20 выпускников, или 9,9%).

2. Построить матрицу

Текст (№)	лицей	ВШЭ	экономика	выпускник	олимпиада
1	**2**	**2**	**2**	**3**	**3**
2	**1**	**0**	**6**	**0**	**2**
3	**0**	**1**	**0**	**1**	**5**

Каждый текст – вектор, представляющий частоту слов.

Wi = (Wi1, Wi2, …, Wim), i – номер текста, m – слово m

Wim – частотность слова m

W1 = (2, 2, 2, 3, 3)

Взвешивание

TF-IDF (**TF** — term frequency, **IDF** — inverse document frequency)

-снижает вес тех слов, которые имеют высокую частотность и встречаются во многих текстах.

-повышает вес тех слов, которые различают отдельные тексты.

Обучающая выборка (training set) – выборка, на основании которой обучается модель.

Тестовая выборка (test set)– выборка, на которой проверяется качество модели.

Обучающая выборка (training set) – выборка, на основании которой обучается модель.

Тестовая выборка (test set)– выборка, на которой проверяется качество модели.

Перекрёстная проверка (cross-validation): разбивка обучающей выборки на n выборок. Обучение на (n-1) выборке, оценка на n-ой выборке.

Алгоритмы

-Метод опорных векторов (support vector machine - SVM)

-Логистическая регрессия

-Наивный байесовский классификатор

-Мультиноминальная логистическая регрессия

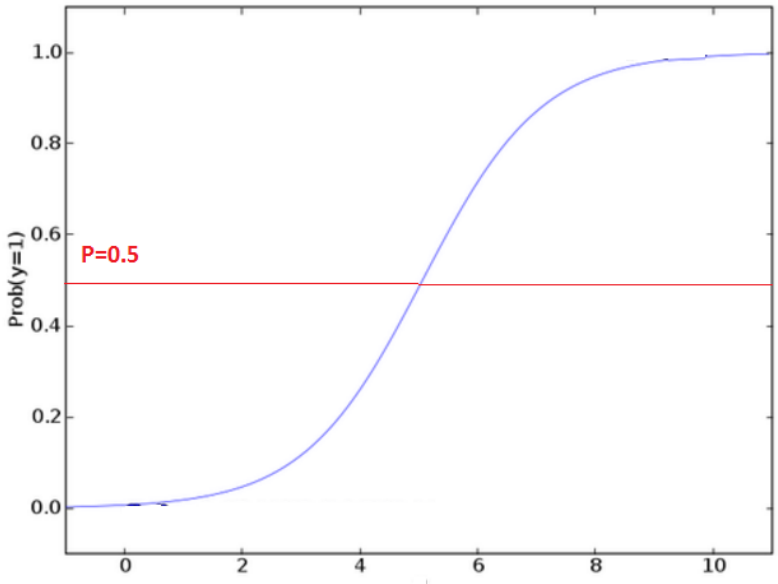
- Деревья решений

И.т.д.

Логистическая регрессия и метод опорных векторов

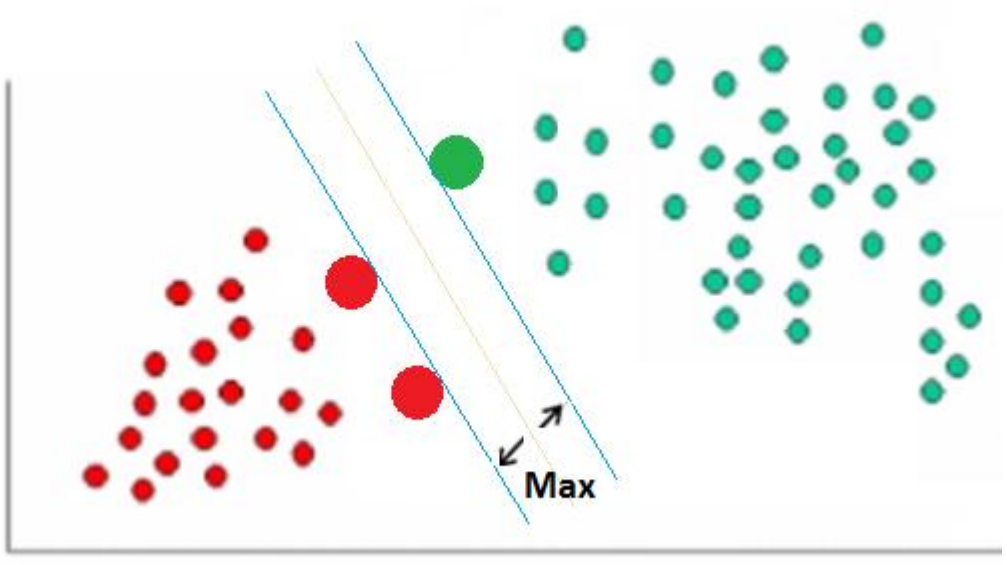
Логистическая регрессия

Log функция:
y=0, если p<0.5
y=1, если p>0.5



Метод опорных векторов

Максимизация зазора



Меры качества

		Экспертная оценка/кодирование	
		Положительная	Отрицательная
Оценка системы	Положительная	**TP**	**FP**
	Отрицательная	**FN**	**TN**

$$\text{Precision} = \frac{TP}{TP + FP}$$

сколько полученных от классификатора ответов являются правильными.

$$\text{Recall} = \frac{TP}{TP + FN}$$

доля правильно классифицированных текстов из всех текстов, которые классификатор отнёс к этой категории.

Переобучение

- Переобучение (overfitting)

- Регуляризация:

L1

L2

Эмпирический кейс

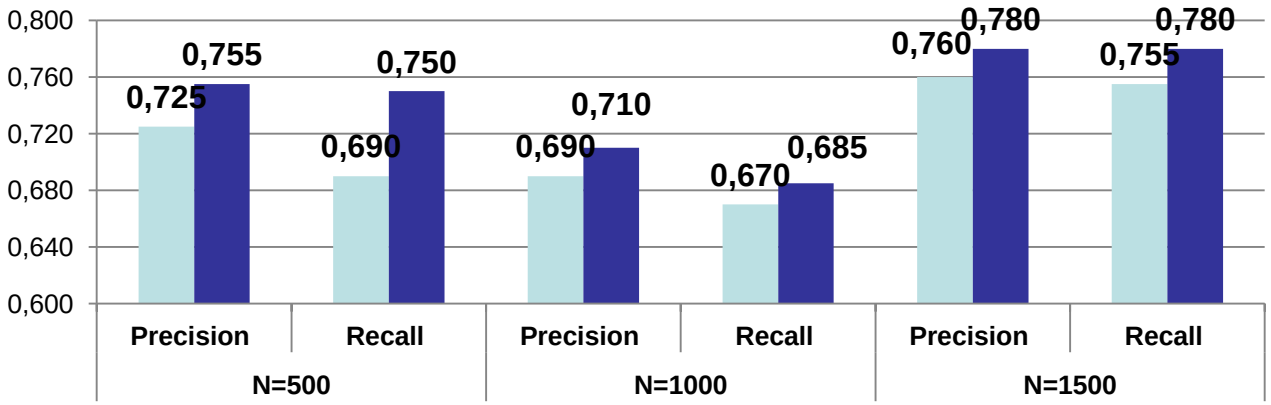
N=26 866 статей
1500 статей закодировано (0: 39%, 1: 61%)

Сравнение двух алгоритмов:
- SVM (метод опорных векторов)
- Логистическая регрессия

Сравнение трёх N:
- N=500
- N=1000
- N=1500

Валидация:
- Без перекрестной валидации
- Перекрестная валидация на 5 сетах
- Перекрестная валидация на 10 сетах

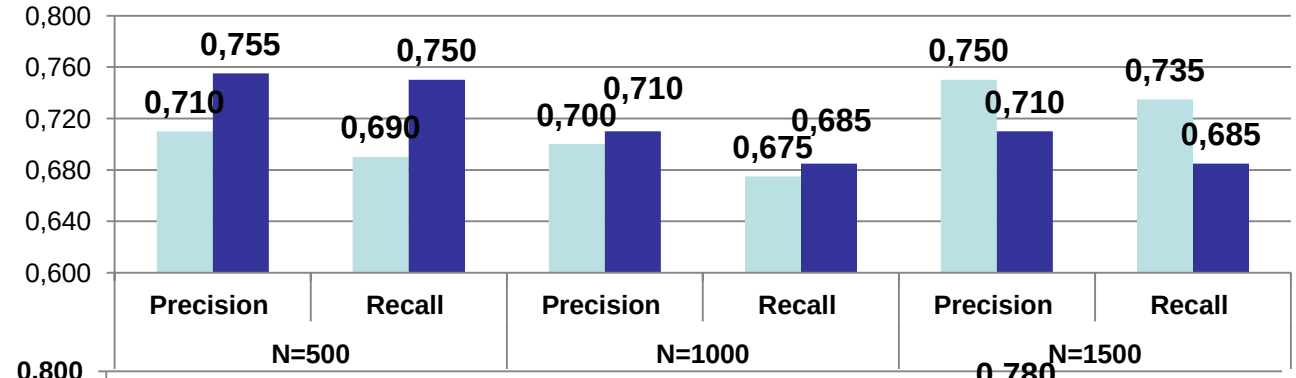
Эмпирический кейс



Без перекрёстной валидации

SVM

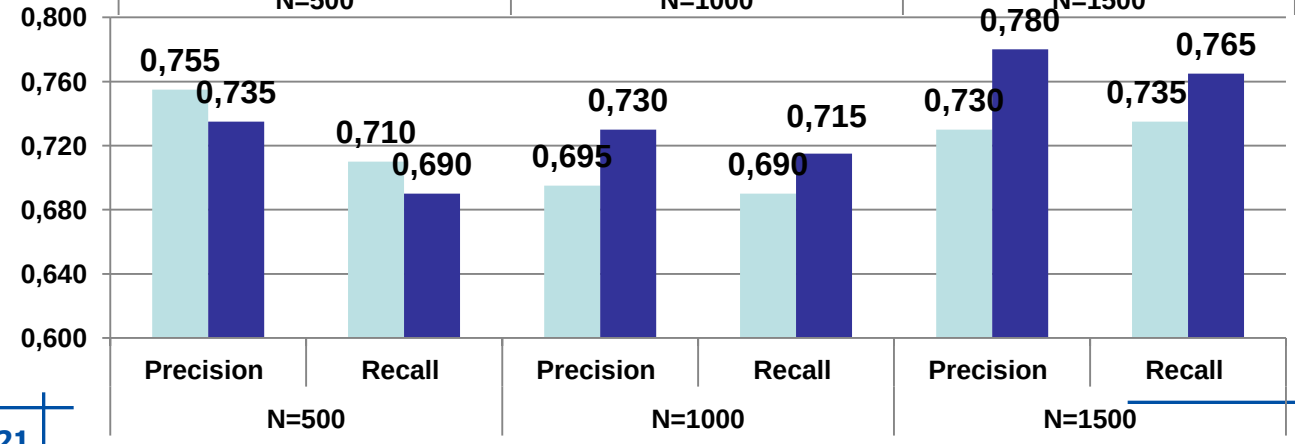
Логистическая регрессия



Валидация на 5 сетах

SVM

Логистическая регрессия

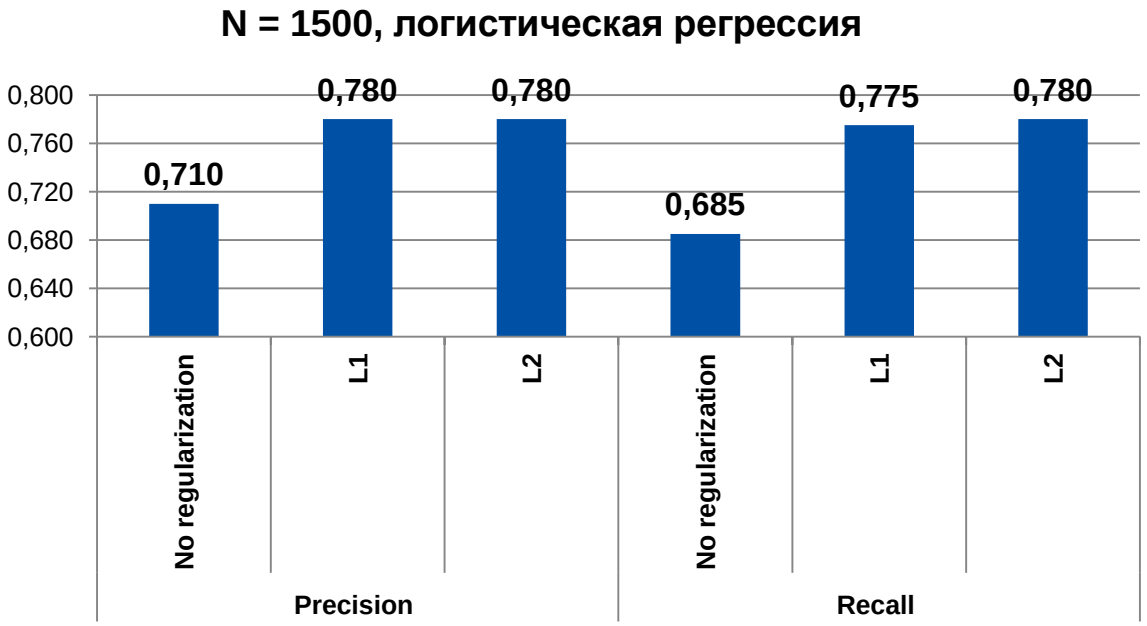


Валидация на 10 сетах

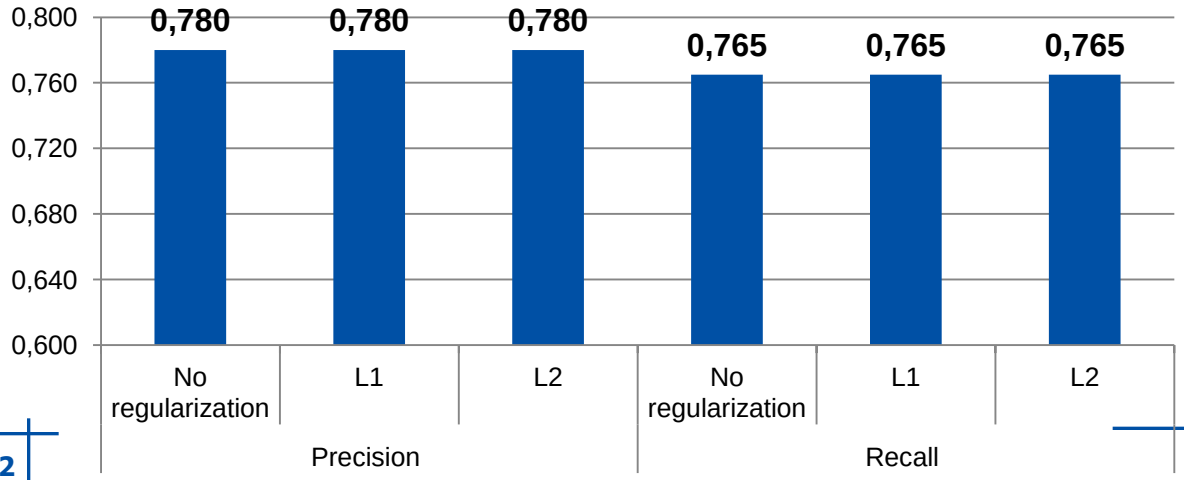
SVM

Логистическая регрессия

Эмпирический кейс



Валидация на 5 сетах



Валидация на 10 сетах

Спасибо за внимание!

Айгуль Мавлетова
amavletova@hse.ru