Три задачи прогноза на основе текстов

Евгений Нижибицкий

22 апреля 2013 г.

1. Предсказание сборов фильмов по их отзывам
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат
2. Предсказание риска по финансовым отчетам
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат
3. Предсказание возраста автора текста
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат

Постановка задачи

Перед тем, как фильм поступает в прокат, критики уже оставляют отзывы об увиденном на предварительном просмотре. Отсюда возникает задача узнать, сколько же получит фильм в прокате? В этой задаче можно рассмотреть много таких признаков, как возрастной рейтинг, жанр, наличие «звёздных» актёров, бюджет, но главную роль всё же будут играть отзывы, созданные профессиональными критиками.

Данные

1351 фильм, в периоде между январём 2005 г. и июнем 2009 г.
Количественные данные:

- `www.metacritic.com`: название, жанр, сценарист, режиссёр, страна производства, основные актёры, дата выхода, возрастной рейтинг (MPAA) и длительность проката.
- `www.the-numbers.com`: бюджет съемок, сборы за первую неделю, количество экранов проката.

Текстовые данные:

- отзывы, собранные с шести обзорных сайтов, наиболее часто упоминавшихся на `www.metacritic.com`. Только отзывы, выпущенные *до* даты выпуска фильма в прокат, были обработаны, чтобы быть уверенными, что никакая информация после стартового уикенда не повлияла на них.

Были получены три типа текстовых признаков:

1. $\mathfrak{n}$-граммы, где $\mathfrak{n}$ рассматривалось от 1 до 3. Стоп-лист состоял из 25 слов. Биграммы и триграммы отфильтровывались только при попадание всех слов в стоп-лист.
2. $\mathfrak{n}$-граммы частей речи — $\mathfrak{n}$ также от 1 до 3. Тэги были получены с помощью приложения из Стэнфорда (Toutanova and Manning, 2000).
3. Зависимость отношений(?). Был использован парсер из Стэнфорда (Klein and Manning, 2003) для обработки отзывов критиков и выделения зависимостей. Получаемые признаки состоят из первой части тройки <отношение, заглавное слово, слово-модификатор>.

В качестве модели рассматривалась линейная регрессия с комбинированной L_1 и L_2 регуляризацией (“elastic net”):

$$\frac{1}{n}\sum_{i=1}^{n}(M_i - \mathbf{w}^\top \mathbf{f_i})^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2 \to \min_{\mathbf{w}}$$

Модель была натренирована на 988 примерах в период 2005–2007 гг., а регуляризационные константы настраивались на периоде январь-агуст 2008 г. Оценка модели проводилась прогнозированием искомого значения для каждого фильма с сентября 2008 г. по июнь 2009 г.

Результат

Модель	Ошибка (MAE)
Базовая (медиана по тренировочным данным)	$7097
Только количественные признаки	$7313
Только текстовые признаки	$6729
Количественные и текстовые признаки	$6725

Как видим, модели, испольщующие текстовую информацию либо её же вкупе с метаданными, выигрывают по сравнению с моделями, использующими только метаданные.

1. Предсказание сборов фильмов по их отзывам
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат

2. Предсказание риска по финансовым отчетам
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат

3. Предсказание возраста автора текста
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат

Постановка задачи

Рассматривалась задача прогнозирования волатильности доходности акций (*stock return volatility*) компании на основе её ежегодного финансого отчета. Волатильность часто использует в финансах как мера риска. Она равна стандартному отклонению доходности акции за конечный промежуток времени. Т.е. акция обладает высокой волатильностью при сильной флуктуации во времени.

Пусть $r_t = \frac{P_t}{P_{t-1}} - 1$ будет доходом с одной акции между закрытием торговых дней $t-1$ и t, где P_t — цена акции во время. закрытия дня t. Тогда измеряемая волатильность за временной период с дня $t-\tau$ по день t равняется отклонению

$$\nu_{[t-\tau,t]} = \sqrt{\sum_{i=0}^{\tau} \frac{(r_{t-i} - \hat{r})^2}{\tau}},$$

где $\hat{r}$ — среднее значение r_t за данный период.

Важно! Прогнозирование волатильности это совсем не то же, что прогнозирование цены акции или её доходности. Вместо предсказывания поведения акции мы предсказываем насколько оно будет стабильно. К данному времени в эконометрики уже пришли к выводу, что прогнозирование цены акции по публичным данным слишком сложно. Это верный аттрибут хорошо функционирующего рынка и краеугольный камень т.н. «гипотезы эффективного рынка» (Fama, 1970). Тем не менее, прогнозирование уровня риска на основе публичных данных ничему не противоречит.

В США все выставленные на биржу компании предоставляют ежегодный отчет, известный как «Форма 10-K». Отчёт обычно включает информацию об истории и внутренней организации компании, её капитале и дочерних предприятиях, а также финансовую информацию. Эти отчеты лежат во всеобщем доступе на официальном сайте Securities Exchange Commission. Используемые в экспериментах данные включали 54379 отчетов, опубликованных 10492 компаниями в период 1996–2006 гг. Каждый отчёт содержит дату публикации, что важно для сопоставления текста с периодом времени, в которые хотим прогнозировать финансовые показатели.

С точки зрения информативности для предсказания, одна из секций отчёта по форме 10-K представляет наибольший интерес — секция 7, известная как «Обсуждение и анализ финансовых условий и результатов деятельности руководством компании», а точнее посекция 7A, «Количественная и качественная информация о рыночных рисках».
Т.к. секция 7 содержит всю наиболее полезную информацию в отчетах, остальная часть отчетов просто отфильтровывалась.

Секция 7 обычно начинается с введения вроде такого (отчет по форме 10-K компании ABC за 1998 г.):

- The following discussion and analysis of ABC's consolidated financial condition and consolidated results of operation should be read in conjunction with ABC's Consolidated Financial Statements and Notes thereto included elsewhere herein. This discussion contains certain **forward-looking statements which involve risks and uncertainties.** ABC's actual results could differ materially from the results expressed in, or implied by, such statements.

В добавок к отчётам, использовались данные по акциям американских компаний с официального сайта «Center for Research in Security Prices (CRSP)». Вычислялись два значения волатильности — за период в 12 месяцев до отчёта ($\nu^{(-12)}$) и после отчёта ($\nu^{(+12)}$).

Для начала из отчёта было получен словарь из M слов на основе тренировочной выборки. Обозначем через $freq(x_j; \mathbf{d})$ количество вхождений j-го слова в словаре в документ $\mathbf{d}$. Тогда далее рассматриваются следующие признаки (все сохраняют разреженность представления документов):

- **TF**: $h_j(\mathbf{d}) = \frac{1}{|\mathbf{d}|} freq(x_j; \mathbf{d}), \forall j \in \{1, \ldots, M\}$.
- **TFIDF**: $h_j(\mathbf{d}) = \frac{1}{|\mathbf{d}|} freq(x_j; \mathbf{d}) \times \log(\frac{N}{|\{\mathbf{d} \mid freq(x_j; \mathbf{d}) > 0\}|})$.
- **LOG1P**: $h_j(\mathbf{d}) = \log(1 + freq(x_j; \mathbf{d}))$.

Т.к. одной из целей было узнать, как добавление информации, полученной из отчётов, улучшает качетсво по сравнению с простым прогнозом на основе истории, был также добавлен $M + 1$-й признак, равный $\log \nu^{(-12)}$.

В качестве модели рассматривалась регрессия с помощью машины опорных векторов (SVR):

$$\frac{C}{N}\sum_{i=1}^{N}\max\left(0,\left|\nu_i - f(\mathbf{d_i};\mathbf{w})\right| - \varepsilon\right) + \frac{1}{2}\left\|\mathbf{w}\right\|_2^2 \to \min_{\mathbf{w}},$$

где $\mathbf{d_i}$ — документы, $\mathbf{w}$ — параметры, C — регуляризационная константа, а ε контролирует ошибку при валидации. Модель находит функцию f, минимизирующую регуляризованный эмпирический риск.

Если h - функция преобразования документов в векторное представление, то в модели SVR функция f примет вид:

$$\mathrm{f}(\mathbf{d}; \mathbf{w}) = \mathrm{h}(\mathbf{d})^{\top}\mathbf{w} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{\mathrm{i}} \mathrm{K}(\mathbf{d}, \mathbf{d_i}),$$

где функция f параметризуется в терминах ядерной функции K с «двойственными» весами α_{i}. K является функцией сходства между двумя документами. При использовании линейного ядра, зависимость между первичными и двойственными весами также линейна:

$$\mathbf{w} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_{\mathrm{i}} \mathrm{h}(\mathbf{d_i}).$$

Для экспериментов использовалась реализация SVM$^{\mathrm{light}}$.

Результат

Ошибка MSE для различных моделей на тестовых данных:

Признаки	2001	2002	2003	2004	2005	2006	среднее
$\nu^{(-12)}$ (baseline)	0.1747	0.1600	0.1873	0.1442	0.1365	0.1463	0.1576
$\nu^{(-12)}$ (SVR with bias)	0.2433	0.4323	0.1869	0.2717	0.3184	5.6778	1.2061
$\nu^{(-12)}$ (SVR without bias)	0.2053	0.1653	0.2051	0.1337	0.1405	0.1517	0.1655
TF	0.2219	0.2571	0.2588	0.2134	0.1850	0.1862	0.2197
TFIDF	0.2033	0.2118	0.2178	0.1660	0.1544	0.1599	0.1842
LOG1P	0.2107	0.2214	0.2040	0.1693	0.1581	0.1715	0.1873
LOG1P, bigrams	0.1968	0.2015	0.1729	0.1500	0.1394	0.1532	0.1667
TF +	0.1885	0.1616	0.1925	0.1230	0.1272	0.1402	0.1541
TFIDF +	0.1919	0.1618	0.1965	0.1246	0.1276	0.1403	0.1557
LOG1P+	0.1846	0.1764	0.1671	0.1309	0.1319	0.1458	0.1542
LOG1P+, bigrams	0.1852	0.1792	0.1599	0.1352	0.1307	0.1448	0.1538

Биграммы с положительным вкладом в волатильность:

- *loss, net loss, year, expenses, going concern, additional.*

Биграммы с отрицательным вкладом в волатильность:

- *dividends, distributions, merger agreement, income, rate.*

1. Предсказание сборов фильмов по их отзывам
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат
2. Предсказание риска по финансовым отчетам
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат
3. Предсказание возраста автора текста
   - Постановка задачи
   - Данные
   - Метод решения
   - Результат

Постановка задачи

Рассматривалась задача прогнозирования возраста автора на основе написанного им сообщения на форуме/оставленной записи в блоге/расшифровки телефонного разговора и т.п.

Данные

Были рассмотрены 3 корпуса текста с разными характеристиками:

- Блоговый корпус
- Корпус Фишера телефонных разговоров
- Форум по раковым заболеваниям

Данные были разделены на тренировку, валидацию и тест.

Содержит сграбленные в 2004 г. блоги с сайта `blogger.com`. Информация о пользователях была предоставлена ими самими в описании профилей. Пользователи были поделены на три возрастные группы с равным количеством представителей обоих полов. Каждый документ в выборке состоит из всех записей конкретного блогера.

Данные

Корпус Фишера телефонных разговоров

Содержит расшифровки телефонных переговоров. Люди произвольным образом были поделены на пары, информация хранилась о каждом испытуемом. Более того, для каждого сеанса связи задавалась тема разговора. Данные собирались в течение года начиная с декабря 2002-го года. Для экспериментов были аггрегированы данные о каждом испытуемом.

Данные были собраны с одного из наиболее активных форумов `http://community.breastcancer.org`. Все записи и профили пользователей были собраны в январе 2011 г. Только малая часть пользователей указывала возраст, поэтому дополнительно был проставлен возраст 200-м пользователям на основе их сообщений, где это можно было сделать точно (к примеру, *I was diagnosed 2 years ago when I was just 38*).

Текстовые признаки:

- Униграммы.
- Униграммы и биграммы из частей речи — n также от 1 до 3. Тэги были получены с помощью приложения из Стэнфорда (Toutanova et al., 2003).
- LIWC (Pennebaker et al., 2001). Эта программа выделяет классы слов, такие как соединительные слова (LIWC-incl: “with”, “and” и другие), слова-причины (LIWC-cause: “because”, “hence” и др.), и стилевые характеристики, такие как, например, процент слов, длиннее шести букв (LIWC-Sixltr).

Т.к. гендерная принадлежность также влияет на связь возраста и речи/текста человека (см. Argamon et al. (2007)), был добавлен соотвествутющий бинарный признак.

В качестве модели рассматривалась линейная регрессия с L_1-регуляризацией(Lasso):

$$\frac{1}{n}\sum_{i=1}^{n}(M_i - \mathbf{w}^\top \mathbf{f_i})^2 + \lambda \left|\left|\mathbf{w}\right|\right|_1 \to \min_{\mathbf{w}}$$

Она минимизирует сумму квадратов отклонений, причем за счет регуляризации достигается разреженность коэффициентов. Регуляризационная константа λ находилась с помощью экспериментов на валидационной выборке.

Чтобы определить, какие признаки являются важными для всех корпусов, а какие являются специфичными, тренировка модели была произведена на объединённых данных с использованием представления признаков, предложенного Daume III (2007). Используя эту модель, признаковое пространство было расширено за счет представления каждого признака как четырёх новых признаков: глобального и трёх специфичных для корпусов. Т.е. для каждого признака f мы получили $\mathrm{f}_{\mathrm{global}}$, $\mathrm{f}_{\mathrm{blogs}}$, $\mathrm{f}_{\mathrm{fisher}}$ и $\mathrm{f}_{\mathrm{cancer}}$.

Для каждого вхождения в документ устанавливается только глобальный и соответствующий признаки. К примеру, для значения признака x_j из корпуса блогов мы получим вектор $(x_j, x_j, 0, 0)$. Т.к. итоговая модель использует L_1-регуляризацию и выделяет небольшую подвыборку признаков, некоторые признаки могут появиться только как глобальные или специфичные для соответствующего корпуса.

Помимо экспериментов с чистой JOINT-моделью исследовалось качество при использовании только выделенных глобальных признаков. Это достигалось применением тех весов для глобальных признаков, которые были получены композиционной моделью, или обучением заново модели на каждом из корпусов, используя только глобальные признаки.

В итоге, были рассмотрены следующие модели:

- INDIV: Модели обучются индивидуально на каждом корпусе.
- JOINT: Модель обучается на совместном корпусе, используя представлениеDaume III.
- JOINT-Global: Используем JOINT-модель, но оставляем только глобальные признаки.
- JOINT-Global-Retrained: Используем глобальные признаки из JOINT-модели, но обучаемся заново на каждом из трёх корпусов.

Результат

В таблице приведены лучшие результаты для каждого из корпусов.

Корпус	Модель	Признаки	Ошибка (MAE)
Blogs	INDIV	все	4.114
Fisher	JOINT	все	6.835
Cancer	JOINT	только униграммы	6.537

Возраст автора — **17** лет:
I can't sleep, but this time I have school tommorow, so I have to try I guess. My parents got all pissed at me today because I forgot how to do the homework [...]. Really mad, I ended it pissing off my mom and [...] NOTHING! Damn, when I'm at my cousin's I have no urge to use the computer like I do here, [...].

Возраст автора — **17** лет:
I can't sleep, but this time I have school tommorow, so I have to try I guess. My parents got all pissed at me today because I forgot how to do the homework [...]. Really mad, I ended it pissing off my mom and [...] NOTHING! Damn, when I'm at my cousin's I have no urge to use the computer like I do here, [...].
Предсказанный возраст — **16.58**.
Этот текст содержит такие характерные признаки, как упоминание школы, родителей а также использование ругательств.

Возраст автора — **19** лет:
Im very young and an athlete and I really do not want to look disfigured, especially when I work so hard to be fit. I know it sounds shallow, but Im young and hope to [...] my husband one day :) [...] My grandmother died of breast cancer at 51, and my mother is currently dealing with a cancerous tumor on her ovaries.

Возраст автора — **19** лет:
Im very young and an athlete and I really do not want to look disfigured, especially when I work so hard to be fit. I know it sounds shallow, but Im young and hope to [...] my husband one day :) [...] My grandmother died of breast cancer at 51, and my mother is currently dealing with a cancerous tumor on her ovaries.
Предсказанный возраст — **35.48**.
При отсутствии явного указания на столь юный возраст, этот текст еще и по стилю является более формальным, чем остальные, что делает его сложным для прогноза.

Возраст автора — **47** лет:
[...]In the weeks leading up to this meeting certain of the managers repeatedly asserted strong positions. [...] their previous (irresponsible yet non-negotiable) opinions [...] Well, today's my first Father's day [...]. Bringing a child into this world is quite a responsibility especially with all the fears and challenges we face.

Возраст автора — **47** лет:
[...]In the weeks leading up to this meeting certain of the managers repeatedly asserted strong positions. [...] their previous (irresponsible yet non-negotiable) opinions [...] Well, today's my first Father's day [...]. Bringing a child into this world is quite a responsibility especially with all the fears and challenges we face.
Предсказанный возраст — **34.42**.
Текст содержит такие характерные признаки, как отсылки к работе и наличию детей. Тем не менее, упоминание отца могло спутать модель.

Возраст автора — **73** года:
T: ah thoughts i'm retired right now
T: i i really can't ah think of anyth- think of i would ah ah change considerably ah i'm i'm very i've been very happily married and i have ah three children and six grandchildren
T: yeah that's right well i i think i would do things more differently fair- fairly recently than a long time ago

Возраст автора — **73** года:
T: ah thoughts i'm retired right now
T: i i really can't ah think of anyth- think of i would ah ah change considerably ah i'm i'm very i've been very happily married and i have ah three children and six grandchildren
T: yeah that's right well i i think i would do things more differently fair- fairly recently than a long time ago
Предсказанный возраст — **73.26**.
Текст содержит отсылки к усталости, наличию внуков и использование характерного“ah”.

Спасибо за внимание!