Линейные модели: жатые чувства, SVM (начнем)

И. Куралёнок, Н. Поваров

Яндекс

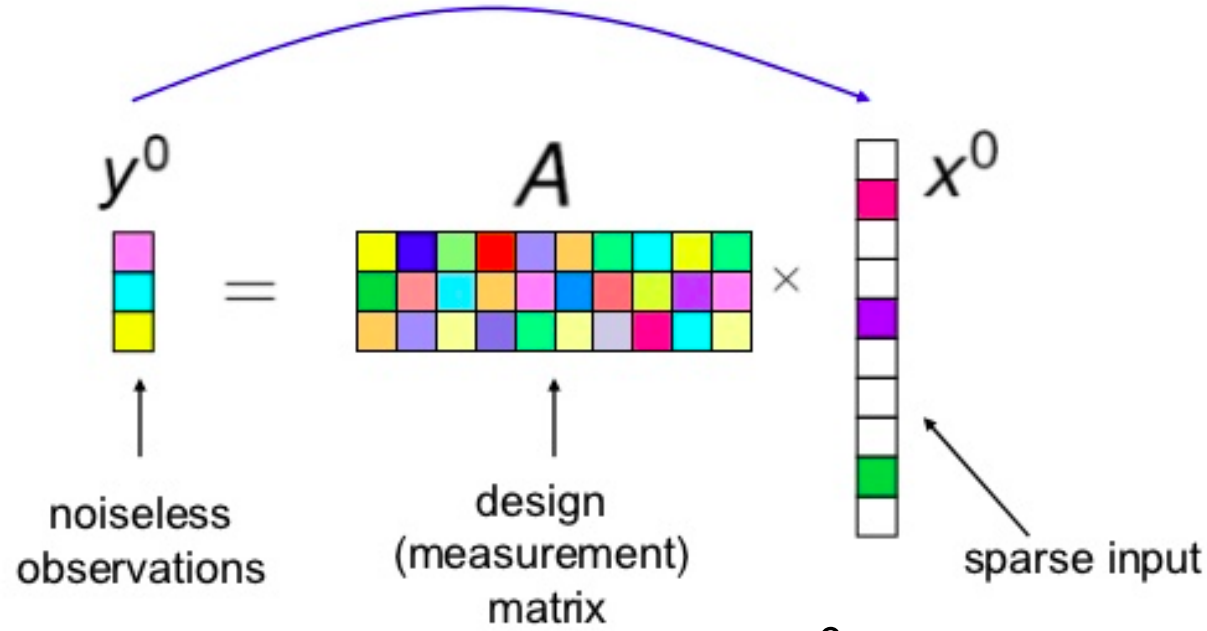
СПб, 2013

План

Пример

Сергей Юрьевич любит смотреть телевизор и рассуждать. Есть мнение, что в основном по телевизору "льют воду". Надо понять как часто надо обращать внимание на то, что проходит на экране, чтобы не упустить "нить".

Пример: постановка задачи



- В телевизоре хотят сказать x^0 (β^*)
- Матрица A (X) — язык передачи
- y^0 (y) — то, что мы видим

$\Rightarrow$ Хотим устроить язык передачи так, чтобы минимизировать количество наблюдений, для восстановления $\hat{\beta}$ как можно ближе к правде β^*

Картинка из Tutorial ICML2010 by Irina Rish & Genady Grabarnik

Сюрприз compressed sensing

$$y = X\beta + \epsilon$$

Если компоненты матрицы X независимые, одинаково распределенные, нормальные, то β можно восстановить точно с большой вероятностью:

- из $O(klog(\frac{m}{k}))$ измерений;
- решив оптимизацию

$$\begin{aligned} &\arg\min_\beta \|\beta\|_1 \\ &\|y - X\beta\| < \epsilon \end{aligned}$$

$\Rightarrow$ где-то мы уже такое видели

Линейная регрессия vs. восстановление сигнала

- Решают одну и ту же задачу
- Одни и те же алгоритмы
- Учиться сложнее:
  - нету влияния на построение матрицы X;
  - в частности нет гарантий на свойства матрицы X;
  - наличие в β большого количество нулей – лишь наше предположение.

Постановка в терминах RFP I

Будем рассматривать множество возможных наблюдений как ось времени, тогда можно рассматривать передачу информации о загаданном β^* как моделирование сигнала через разложение в Фурье. При этом, для простоты, будем считать, что количество возможных наблюдений совпадает с размерностью вектора β^*, в этом случае мы можем рассматривать преобразование как линейную систему DFT:

$$\begin{aligned} y &= \mathcal{F}\beta^* \\ &= \sum_{t=1}^{n} \beta_t e^{\frac{-2\pi i \omega t}{n}} \end{aligned}$$

Возвращаясь к примеру, для Сергея Юрьевича ситуация выглядит как-то так:

$$\beta^* = \mathcal{F}^{-1}\mathcal{F}\beta^* = \frac{1}{n}\mathcal{F}^*\mathcal{F}\beta^*$$

План

Постановка в терминах RFP II

$$\arg\min_\beta \|\beta\|_1$$
$$\|y - X\beta\| < \epsilon$$

В новых обозначениях:

$$\arg\min_\beta \|\beta\|_1$$
$$\|(\mathcal{F}\beta)_\Omega - (\mathcal{F}\beta^*)_\Omega\| < \epsilon$$

LASSO для восстановления сигнала

Для начала решим задачу в которой наблюдения точные:

$$y = (\mathcal{F}\beta^*)_k, k \in \Omega$$

При этом будем решать

$$\begin{aligned} &\arg\min_\beta \|\beta\|_1 \\ &(\mathcal{F}\beta)_k = (\mathcal{F}\beta^*)_k, k \in \Omega \end{aligned}$$

с равными размерностями β^* и $\mathcal{F}\beta^*$.

Теорема о качестве восстановленного сигнала для RFP

Theorem (Candes et al. (2006))

$\beta \in \mathbb{C}^n, |\{i \in \mathbb{Z}_n | \beta_i^* \neq 0\}| = S$

$\Omega \subset \mathbb{Z}_n$ — *одно из равновероятных множеств размера* n

зафиксируем точность B

$\Rightarrow$ *с вероятностью* $P \geq 1 - O(n^{-B})$ *мы можем точно восстановить* β^*, *если:*

$$|\Omega| \geq C_B' S \log n$$

где $C_B' \simeq 23(B+1)$

Выводы из теоремы

- Теорема рассказывает о свойствах случайной DFT проекции
- Загаданный вектор x может быть восстановлен:
  - с высокой вероятностью
  - используя LASSO
  - количество наблюдений пропорционально количеству ненулей в “загаданном” сигнале

Упрощение рандома

В теореме Ω равномерно распределена по всем множествам размера n. Такое сложно генерировать. Значительно проще $\Omega' : \forall j \in \mathbb{Z}_n, P(j \in \Omega) = \tau$.
$\Rightarrow$ Для таких проекций вероятность восстановить сигнал примерно такая же.

Стабильно ли решение?

Интересны два вида “стабильности”:

стабильность: маленькие изменения в решении при малом изменении в наблюдениях (изменения в загаданном);

робастность: устойчивость к шуму в данных (неточно померяли отлик x).

Если мы уже решили проблему построения T, то решение стабильно:

$$\hat{\beta} = (\mathcal{F}^*_{T,\Omega}\mathcal{F}_{T,\Omega})^{-1}\mathcal{F}^*_{T,\Omega}y$$

Из доказательства теоремы о восстановлении сигнала $\mathcal{F}^*_{T,\Omega}\mathcal{F}_{T,\Omega} > \delta E$ с высокой вероятностью при условии на Ω. А вот с робастностью все сложнее...

А что же с произвольно построенным X?

Пока Сергей Юрьевич получал закодированный в Фурье сигнал и раскодировал его обратным Фурье. А что, если кодировани и раскодирование сигнала происходит как-то иначе. Положим, что так:

$$\beta^* = \mathbf{\Phi}^{-1}\mathbf{\Phi}\beta^* = \mathbf{\Psi}\mathbf{\Phi}\beta^*$$

Будем рассматривать ортонормированные $\mathbf{\Phi}, \mathbf{\Psi}$

Когерентность базисов

Definition

Для пары ортонормированных базисов назовем

$$\mu(\Phi, \Psi) = \sqrt{n} \max_{i,j} |(\phi_i, \psi_j)|$$

когерентностью.

- Заметим, что $1 \leq \mu(\Phi, \Psi) \leq \sqrt{n}$
- В случае Фурье получается экстремально хороший случай: $\mu(DFT, IDFT) = 1$

Теорема о качестве восстановленного сигнала для произвольных базисов

Theorem (Candes and Romberg (2006))

Для фиксированной $\delta > 0$ *и* $x \in \mathbb{R}^n$, $|\{i|\beta_i^* \neq 0\}| < S$. *Выберем* Ω *точек для наблюдения равномерно из* $\mathbb{Z}_n$ *без повторений. Если*

$$|\Omega| \geq C\mu^2(\Phi, \Psi)S \log \frac{n}{\delta}$$

тогда решение LASSO:

$$\begin{aligned} &\arg\min_{\beta \in \mathbb{R}^n} \|\beta\|_1 \\ &(\Phi\beta)_\Omega = (\Psi x)_\Omega \end{aligned}$$

восстановит x *с вероятностью* $1 - \delta$

Возвращаемся к случаю шумных наблюдений

Воспользовавшись построенной теорией для точных наблюдений, введем ряд дополнительных ограничений:

1. Вводим ограничение на модельную матрицу (Restricted Isomenry Property):

$$\exists \delta(S = |\{i|x \neq 0\}|) :$$
$$(1 - \delta(S))\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta(S))\|x\|_2$$

2. В введенных условиях получаем ограничение на робастность в рамках восстановления сигнала

3. Переходим от когерентности к условиям на собственные числа модельной матрицы

LASSO persistency theorem

Во введенных условиях оказывается, что (LASSO persistency theorem, Bickel et al., 2009):

$$\|\hat{\beta} - \beta^*\| \leq O\left(\sqrt{\frac{\log n}{m}}\right)$$

Сравним полученное с условиями на несмещенное решение, где мы могли легко убежать бесконечно далеко от заданного β^*.

Что мы узнали про CS

1. Можно ставить задачу по восстановлению сигнала
2. Для решения задачи нам понадобится рандомно выбирать точки наблюдения
3. Оказывается, что решать подобные задачи нужно тем же самым LASSO
4. Эффективность решения зависит от того, как построить “язык передачи информации”
5. Одним из самых хороших универсальных языков (с минимально возможной когерентностью) является DFT/IDFT
6. С помощью механизма CS можно доказать устойчивость решения LASSO

План

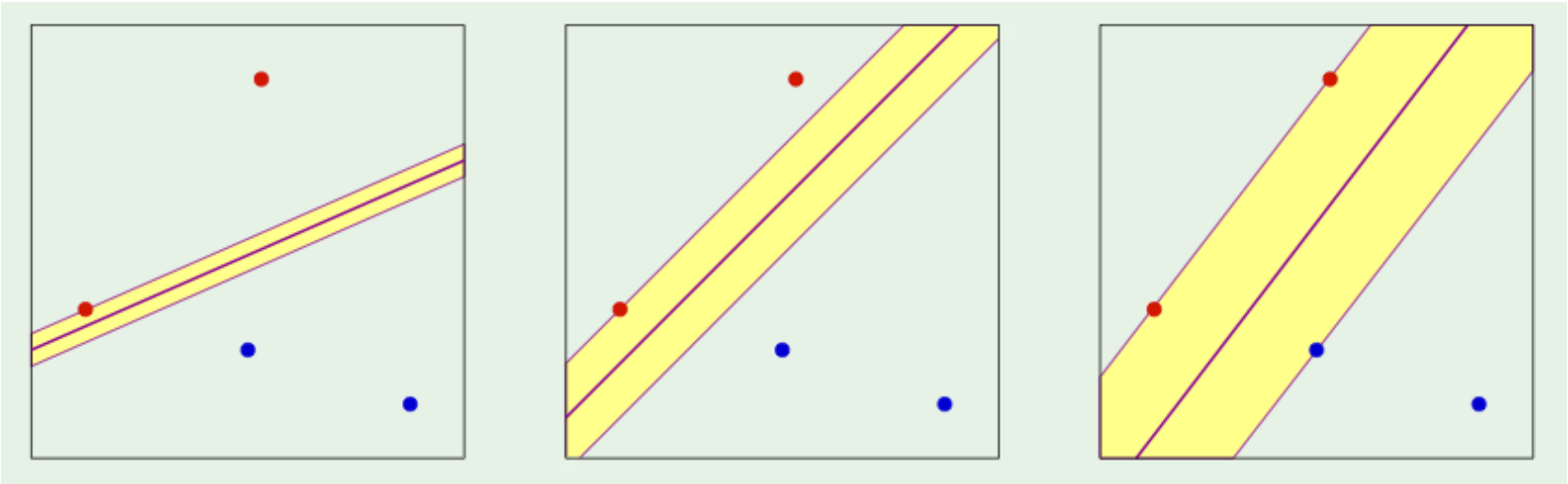
1. Постановка задачи восстановления сигнала
   - Пример
   - Разложение сигнала в Фурье и постановка в нахождении коэффициентов
2. LASSO для восстановления сигнала
   - Теорема о качестве восстановленного сигнала (Candes et al. 2006)
   - Стабильность решения: RIP, RRfND (Candes et al. 2006)
   - LASSO persistency theorem (Bickel et al., 2009)
3. Support vector machines
   - Идея метода
   - Коэффициенты Лагранжа для решения задачи про максимальное расстояние
4. Домашнее задание

SVM(воспоминания о былом)

- Последний из линейных методов, который мы рассмотрим подробно.
- Rocket science до конца 90-х, по крайней мере в задачах классификации.

SVM на пальцах

- Максимальный зазор.
- Нелинейные преобразования.



Мысли вслух

- Почему большой зазор это хорошо?
- Какая β максимизирует зазор?

Найдем ширину “зазора”: геометрия

Есть две параллельные плоскости:

$$\begin{cases} \beta^T x = a \\ \beta^T x = b \end{cases}$$

проведем прямую, перпендикулярную этой плоскости: $y = \|\beta\| \frac{\beta}{\|\beta\|} t$. Пересечет она наши плоскости вот так:

$$\begin{cases} \beta^T (\|\beta\| \frac{\beta}{\|\beta\|} t_a) = a \\ \beta^T (\|\beta\| \frac{\beta}{\|\beta\|} t_b) = b \end{cases}$$

$$\begin{cases} t_a = \frac{a}{\|\beta\|} \\ t_b = \frac{b}{\|\beta\|} \end{cases}$$

тогда расстояние по полученной прямой: $|t_a - t_b| = \frac{|a-b|}{\|\beta\|}$

Найдем ширину "зазора": мат. анализ

Решим оптимизацией:

$$\min \frac{1}{2}\|x-y\|^2$$

$$\begin{cases} \beta^T x = a \\ \beta^T y = b \end{cases}$$

Перейдем к коэффициентам Лагранжа:

$$\min \frac{1}{2}\|x-y\|^2 + \lambda_1(\beta^T x - a) + \lambda_2(\beta^T y - b)$$

Найдем нули производных по всем переменным:

$$\begin{cases} \beta^T x = a \\ \beta^T y = b \\ x - y + \lambda_1 \beta = 0 \\ x - y + \lambda_2 \beta = 0 \end{cases} \quad \begin{cases} \beta^T (x-y) = a - b \\ \lambda_1 = \lambda_2 \\ \|\beta\|\lambda_1 = b - a \end{cases} \quad \begin{cases} \lambda_1 = \lambda_2 = \frac{b-a}{\|\beta\|^2} \\ x - y = \frac{b-a}{\|\beta\|^2}\|\beta\|\left(\frac{\beta}{\|\beta\|}\right) \end{cases}$$

Возвращаясь к SVM

Теперь мы знаем что оптимизировать. Отнормируем разделяющие плоскости так:

$$\begin{cases} \beta^T x = b - 1 \\ \beta^T x = b + 1 \end{cases}$$

В этих терминах нас $|a - b|$ фиксированы и оптимизировать мы будем только β:

$$\arg\min \frac{\|\beta\|}{2}$$

Вот в таких условиях ($y_i \in \{-1, 1\}$):

$$y_i(\beta^T x_i - b) \geq 1$$

По методу Лагранжа

По теореме Куна-Таккера:

$$\mathcal{L} = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{m} \lambda_i (y_i(\beta x_i - \beta_0) - 1), \lambda_i \geq 0$$

$$\begin{cases} -\mathcal{L} = -\sum_{i=1}^{m} \lambda_i + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i x_j) \\ \lambda_i \geq 0 \\ \sum_{i=1}^{m} \lambda_i y_i = 0 \end{cases}$$

Тогда:

$$\beta = \sum_{i=1}^{m} \lambda_i y_i x_i$$
$$\beta_0 = \beta x_i - y_i, \lambda_i > 0$$

Чем стало легче?

- Адовые условия сменились простым $\lambda_i > 0$
- У нас получился квадрат количества точек
- Интересны только (x_i, x_j) с которыми мы можем играться (kernel trick)!

План

Результаты ДЗ про придумать таргет

1. c8a9ac - 1
2. 1f7d2b - 1
3. 4da958 - 2
4. 64d24a - 2
5. d3905c - 2
6. 2b2904 - 2
7. 6af9f9 - 3
8. 4afcbe - 3
9. dcd1b7 - 3
10. d1393f - 3
11. b764ae - 4
12. 5266fc - 4
13. 2dd08e - 4
14. 326690 - 4
15. 620441 - 4
16. e7d20b - 4
17. 2f1218 - 4
18. 9b423e - 4
19. 7a3ccc - 5
20. 93203b - 6

Результаты ДЗ (комментарий)

1. Про диагностику насморка - всё просто и решили !!почти!! все
2. Про диагностику рака - многие вспомнили про бесконечные штрафы, но про то, что лечение от рака для здоровья небесплатно не вспомнил никто
3. Про кризисное состояние - только некоторые поняли, что в кризисном состоянии некоторые диагнозы не имеют смысла, так как неизлечимы
4. Про пребывание в больнице - у всех простое и неинтересное решение

Результаты ДЗ (советы)

1. Надо помнить про бесконечные штрафы
2. Надо помнить про эксплуатацию, а не только формально считать число ошибок
3. Кроме точности/полноты/аккуратности у которых есть проблема в случае перекошенной выборки есть такие штуки, как чувствительность/специфичность/AUC
4. Целевая функция != факторы и целевая функция != решающая функция

Домашнее задание

- так как svm сегодня рассказан не полностью, то домашнее задание по нему будет на следующей лекции;
- хинт - задание будет по svm, датасет будет тот же;
- дедлайн будет - 28 ноября.