Принцип максимизации зазора в структурном обучении

Дмитрий Ветров
МГУ ВМК
VetrovD@yandex.ru

1 Введение

Одним из наиболее актуальных вопросов в теории графических моделей является вопрос выбора наилучшей вероятностной модели, описывающей имеющиеся в распоряжении пользователя данные. Концептуально, задача может быть разбита на две подзадачи

- Выбор структуры графа, задающего марковскую или байесовскую сеть, т.е. выбор способа факторизации совместного распределения переменных модели;
- Настройка параметров совместного распределения.

Ниже будет рассмотрено несколько подходов к решению второй задачи, являющейся обобщением классической задачи оценивания параметров распределения. Общих методов для решения первой задачи пока не придумано.

Предполагается, что читатель уже знаком с основными понятиями теории марковских случаяных полей и методами вывода в них.

2 Классический метод опорных векторов

Сначала рассмотрим стандартную задачу классификации на два класса. Пусть нам задана обучающая выборка $(X, \vec{t}) = \{(\vec{x}_i, t_i)\}_{i=1}^n$, где $\vec{x}_i \in \mathbb{R}^d$ — вектор наблюдаемых признаков, описывающий i-ый объект обучающей выбобрки, а $t_i \in \{-1, +1\}$ — метка класса, к которому данный объект принадлежит. Мы хотим по ней построить решающее правило, относящее произвольный объект с признаками $\vec{x}$ к одному из классов по следующему правилу

$$\tilde{t}(\vec{x}) = \text{sign}(\vec{w}^T\vec{x} + b),$$

найдя параметры $\vec{w}$ и b, являющиеся решением следующей задачи условной оптимизации

$$\frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{\vec{w},b,\vec{\xi}} \tag{1}$$

$$\text{s.t.}\;\; t_i(\vec{w}^T\vec{x}_i + b) \geq 1 - \xi_i, \quad \forall i = 1,\ldots,n; \tag{2}$$

$$\xi_i \geq 0 \tag{3}$$

Смысл задачи довольно прост. Мы хотим, чтобы 1) объекты обучающей выборбки относились обученным классификатором к правильному классу, вводя неотрицательный штраф ξ_i за ошибку классификации i-го объекта обучающей выборки; и 2) чтобы при этом веса линейной комбинации признаков принимали не слишком большие значения, вводя штраф на значение квадрата L2-нормы весов. Параметр C определяет соотношение между штрафом за ошибочно классифицированные объекты и штрафом за большие значения нормы $\vec{w}$, т.е. играет роль коэффициента регуляризации.[1]

Введем обозначения $L(X,\vec{t},\vec{w}) = \sum_{i=1}^{n} t_i(\vec{w}^T\vec{x}_i + b)$ и $\xi = 2\sum_{i=1}^{n}\xi_i$. Тогда условие (2) можно переписать в виде

$$L(X,\vec{t},\vec{w}) \geq L(X,\hat{\vec{t}},\vec{w}) + \Delta(\vec{t},\hat{\vec{t}}) - \xi, \quad \forall\hat{\vec{t}} \in \{-1,+1\}^n, \tag{4}$$

где $\Delta(\vec{t},\hat{\vec{t}})$ — манхэттенское расстояние[2] между векторами $\vec{t}$ и $\hat{\vec{t}}$. При выводе выражения (4) мы воспользовались фактом $t(\vec{w}^T\vec{x}+b) = -t'(\vec{w}^T\vec{x}+b)$ при $t \neq t'$. Заметим, что выражение (4) представляет собой набор из 2^n условий, что делает его использование в традиционном методе опорных векторов непрактичным. Тем не менее, оно открывает возможности для обобщения этого подхода на случай, когда обучающая выборка не может быть представлена в виде набора независимых объектов.

3 Многоклассовый метод опорных векторов

Начнем постепенно усложнять задачу. Допустим, что теперь у нас имеется K классов, к одному из которых надо отнести объект. Индексы классов объектов обучающей выборки будем задавать матрицей $T = (\vec{t}_1,\ldots,\vec{t}_n)$, где $\vec{t}_i \in \mathcal{T} = \{\vec{t} \in \{0,1\}^K | \sum_{j=1}^{K} t_{ij} = 1\}$ — бинарный вектор, в котором присутствует лишь одна единица в той позиции, номер которой является индексом класса i-го объекта. Признаковое описание остается тем же, что и в

[1] Более подробно о методе опорных векторов см., например, курс лекций К.В. Воронцова "Математические методы распознавания образов"

[2] Сумма модулей разностей сооответствующих компонент двух векторов.

предыдущем случае. Решающее правило будет иметь вид[3]

$$\tilde{t_j}(\vec{x}) = \begin{cases} 1, & j = \arg\max_j \vec{w}_j^T \vec{x}; \\ 0, & \text{иначе.} \end{cases}$$

В процессе обучения необходимо настроить параметры $W = (\vec{w}_1, \ldots, \vec{w}_K)$, представляющие собой наборы коэффициентов линейной комбинации признаков для каждого класса. Как и ранее, настройка осуществляется исходя из требования минимизации ошибки на обучающей выборке и минимизации нормы весов:

$$\frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{W,\vec{\xi}} \tag{5}$$

$$\text{s.t. } \vec{t_i}^T W^T \vec{x}_i \geq \vec{t} W^T \vec{x}_i + \Delta(\vec{t}, \vec{t}_i) - \xi_i, \quad \forall i = 1, \ldots, n; \ \forall \vec{t} \in \mathcal{T} \tag{6}$$

$$\xi_i \geq 0. \tag{7}$$

Легко показать, что при $K = 2$ эту оптимизационную задачу можно привести к классическому методу опорных векторов. Введя обозначения $\xi = \sum_{i=1}^{n}\xi_n$ и $L(X, T, W) = \sum_{i=1}^{n}\vec{t}_i W^T \vec{x}_i$, можно заменить условия (6) на эквивалентные им

$$L(X, T, W) \geq L(X, \hat{T}, W) + \Delta(T, \hat{T}) - \xi, \quad \forall \hat{T} \in \mathcal{T}^n.$$

Как и в случае двухклассовой задачи, работать с такой формулировкой условий менее удобно, т.к. вместо nK неравенств их стало K^n.

4 Идея структурного метода опорных векторов

В отличие от классической задачи машинного обучения, в структурном обучении предполагается, что данные представимы в виде объектов, между которыми существуют внутренние зависимости, не позволяющие нам принимать решение о классификации данного объекта в отрыве от остальных объектов выборки.[4] Идея обобщения метода опорных векторов на случай структурного обучения заключается в следующем. Пусть имеется обучающая информация, состоящая из набора наблюдаемых переменных X и дискретных

[3]Здесь и далее будем опускать свбодный член b для простоты выкладок.

[4]Простейшим примером такой задачи может являться задача определения победителя выборов в одномандатном избирательном округе по набору характеристик каждого кандидата. Очевидно, что мы не можем решать эту задачу независимо для каждого кандидата, т.к. у нас есть дополнительное структурное условие, что победителем может быть один и только один из кандидатов.

скрытых переменных T, которые для обучающей выборки известны. Структура взаимозависимостей между переменными предполагается известной. Для такой модели вводится параметрически заданная модель $E(X,T,W)$, как правило, являющаяся линейной функцией от настраиваемых весов W. Дальше по аналогии с многоклассовым методом опорных векторов решается задача

$$\frac{1}{2}\|W\|^2 + C\xi \to \min_{W,\xi} \tag{8}$$

$$\text{s.t.} \quad L(X,T,W) \geq L(X,\hat{T},W) + \Delta(T,\hat{T}) - \xi, \quad \forall \hat{T}; \tag{9}$$

$$\xi \geq 0. \tag{10}$$

Поскольку в (9) обычно содержится экспоненциальное число неравенств, дальше применяется стандартный математический прием и условие (9) заменяется на

$$L(X,T,W) \geq \max_{\hat{T}} \Big(L(X,\hat{T},W) + \Delta(T,\hat{T}) \Big) - \xi. \tag{11}$$

Во многих случаях возникающую задачу оптимизации удается решить точно или приближенно за полиномиальное время. Заметим, что формализм структурного метода опорных векторов является весьма общим и применим не только для настройки параметров марковских случайных полей. Но далее мы будем рассматривать именно этот практически важный случай.

Обозначим граф, задающий марковское случайное поле $G = (V,E)$. В дальнейшем будем полагать, что в поле присутствуют потенциалы первого (унарные) и второго (парные) порядка[5] относительно скрытых переменных T. Зависимости между наблюдаемыми переменными нас не интересуют, т.к. значения каждой из них мы наблюдаем непосредственно. Таким образом, с точностью до константы логарифм совместного распределения переменных может быть записан в следующем виде

$$L(T) = \sum_{i \in V} \phi_i(\vec{t}_i) + \sum_{(i,j) \in E} \phi_{ij}(\vec{t}_i, \vec{t}_j),$$

где $\vec{t}_i \in \mathcal{T}$ — бинарный вектор, задающий метку класса i-ой вершины графа G. Будем считать, что для каждой вершины графа задан вектор признаков (наблюдаемых переменных) $\vec{x}' \in \mathbb{R}^{d_1}$, а для каждого ребра — вектор признаков $\vec{x}'' \in \mathbb{R}^{d_2}$. Введем такую

[5]Сколько-нибудь практических методов для настройки потенциалов более высоких порядков (т.н. higher order potentials) в настоящее время не создано, хотя работы над этим ведутся в разных исследовательских группах в мире.

параметризацию потенциалов

$$\phi_i(\vec{t_i}) = \sum_{k=1}^{K} \sum_{s=1}^{d_1} t_{ik} w'_{sk} x'_{is}, \tag{12}$$

$$\phi_{ij}(\vec{t_i}, \vec{t_j}) = \sum_{k,l=1}^{K} \sum_{s=1}^{d_2} t_{ik} t_{jl} w''_{s,kl} x''_{ij,s}, \tag{13}$$

где $W' \in \mathbb{R}^{d_1 \times K}$, $W'' \in \mathbb{R}^{d_2 \times K \times K}$ — веса признаков вершин и ребер, соответственно. Их значения можно найти решив задачу (8)–(10).

5 Метод отсекающей плоскости в структурном обучении

Несмотря на кажущуюся простоту ограничения (11), которое одно заменило экспоненциальное количество ограничений вида (9), решение задачи (8) с таким ограничением затруднительно, т.к. к такой формулировке не могут быть напрямую применены стандартные методы условной оптимизации. Одним из возможных подходов является т.н. метод отсекающей плоскости (cutting plane), работающий следующим образом. Формируется множество Υ т.н. активных ограничений. На первой итерации $\Upsilon = \emptyset$. На каждой итерации решается оптимизационная задача квадратичного программирования (8) с ограничениями (10) и

$$L(X, T, W) \geq L(X, \hat{T}, W) + \Delta(T, \hat{T}) - \xi, \quad \forall \hat{T} \in \Upsilon.$$

Обозначим ее решение (W_*, ξ_*). Затем находится наиболее нарушаемое ограничение

$$T_* = \arg\max_{\hat{T}} L(X, \hat{T}, W_*) + \Delta(T, \hat{T}) = \arg\max_{\hat{T}} D(\hat{T}), \quad \hat{T} \in \mathcal{T}^{|V|}. \tag{14}$$

Если $\max_{\hat{T}} D(\hat{T}) > L(X, T, W_*) + \xi_*$, то T_* добавляется в множество активных ограничений Υ и процесс переходит на следующую итерацию.[6] Если же неравенство не выполнено, это означает, что при данных значениях (W_*, ξ_*) все ограничения удовлетворены и найден оптимум.

[6] Очевидно, что $D(\hat{T}) \leq L(X, T, W_*) + \xi_*$ для всех $\hat{T} \in \Upsilon$.

Как говорилось выше, в качестве функции $\Delta(T, \hat{T})$ используется манхэттенское расстояния, которое может быть переписано в виде[7]

$$\Delta(T, \hat{T}) = \sum_{i \in V} \sum_{j=1}^{K} |t_{ij} - \hat{t}_{ij}| = 2 \sum_{i \in V} \sum_{j=1}^{K} \hat{t}_{ij} [\hat{t}_{ij} \neq t_{ij}].$$

Таким образом, функцию $D(\hat{T})$ можно представить в виде парно-сепарабельной функции

$$D(\hat{T}) = \sum_{i \in V} \sum_{k=1}^{K} \theta_{ik} \hat{t}_{ik} + \sum_{(i,j) \in E} \sum_{k,l=1}^{K} \theta_{ij,kl} \hat{t}_{ik} \hat{t}_{jl},$$

представляющей собой логарифм правдоподобия некоторого марковского поля. В тех случаях, когда задача поиска максимума $D(\hat{T})$ может быть решена точно (например, когда G дерево или $K = 2$ и парные потенциалы обладают свойством субмодулярности) алгоритм отсекающй плоскости позволяет эффективно обучать структурный метод опорных векторов. В случае, когда точная максимизация $D(\hat{T})$ невозможна, используют приближенные схемы, основанные на поиске нижней (жадный подход, undergenerating framework) или верхней (релаксационный подход, overgenerating framework) оценок на значение $\max D(\hat{T})$.

Жадный подход более прост в реализации, т.к. для него достаточно воспользоваться любым приближенным методом вывода в марковских полях, возвращающим некоторую разметку, близкую к оптимальной, например, TRW, MCMC, вариационным подходом, α-расширением (последний при некоторых дополнительных ограничениях), и др. При таком подходе множество, на котором ищется решение задачи (8), **расширяется** относительно допустимого и полученное решение может не удовлетворять условиям (9).

В релаксационном подходе мы получаем верхнюю оценку на значение $\max D(\hat{T})$. Таким образом, множество, на котором ищется решение, **сужается**, а полученное решение заведомо удовлетворяет всем условиям (9). Для получения верхней оценки можно воспользоваться ЛП-релаксацией, либо решением двойственных задач в методах TRW, SMD, возвращающим нецелостную разметку. Результаты последних исследований показывают, что релаксационный подход позволяет получать более точные решающие правила.

6 Метод двойственной задачи в структурном обучении

Рассмотрим другой возможный способ решения задачи (8) при ограничениях (9),(10). Он заключается в замене оптимизационной задачи (14) на двойственную и выносе ее во внешний опимизируемый функционал. Напомним, что двойственные задачи возникают при

[7]Здесь квадратные скобки выдают значение 1, если выражение внутри них верно и ноль в противном случае.

использовании ряда методов приближенного байесовского вывода, использующих двойственное разложение, например, TRW, SMD. Идея заключается в переходе от функционала $D(\hat{T})$ прямой задачи к функионалу $F(\Lambda)$ двойственной задачи.[8] Из теории двойственности известно, что

$$\max D(\hat{T}) \leq \min F(\Lambda).$$

Тогда задачу (8)–(10) можно преобразовать к виду

$$\frac{1}{2}\|W\|^2 + C\xi \to \min_{W,\xi,\Lambda} \tag{15}$$

$$\text{s.t.} L(X,T,W) \geq F(\Lambda) - \xi; \tag{16}$$

$$\xi \geq 0. \tag{17}$$

Основным достоинством такого подхода является отсутствие вложенной оптимизационной задачи, достигнутое за счет увеличения размерности задачи из-за введения новых переменных Λ. Обратите внимание, что такой переход стал возможен только после того как в условиях возникла задача минимизации вместо задачи максимизации. В самом деле, условие

$$L(X,T,W) \geq \max D(\hat{T})$$

заменяется на условие

$$L(X,T,W) \geq \min F(\Lambda).$$

Чем ниже значение $\min F(\Lambda)$, тем шире область допустимых значений W. Но чем она шире, тем ниже минимум функции (8), взятый на этом множестве. Такая положительная связь позволяет минимум по Λ занести во внешнюю оптимизационную задачу. Это довольно общий прием в математических задачах. Заметим, что для применимости такого подхода принципиально, чтобы $F(\Lambda)$ была оценкой сверху на максимум $D(\hat{T})$[9] и, следовательно, такой подход эквивалентен релаксационному методу отсекающей гиперплоскости.

7 Другие методы структурного обучения

Кратко укажем альтернативные подходы к настройке параметров марковских полей.

[8] Здесь Λ — вектор двойственных переменных (множителей Лагранжа), соответствующих релаксируемым ограничениям (см. материалы курса по двойственным задачам).

[9] Иначе минимизация $F(\Lambda)$ вместо приближения к максимуму $D(\hat{T})$ будет приводить к удалению от него.

Принцип максимума правдоподобия. Наиболее естественный путь настройки параметров вероятностного распределения по данным. Параметризовав совместное распределение всех переменных модели $p(T) = p(T|X,W) = \frac{1}{Z(X,W)}\exp(L(X,T,W)$ вектором весов W можно оценить их методом максимального правдоподобия

$$W_{ML} = \arg\max_W p(T|X,W).$$

В случае ацикличных графов (напрмиер, в скрытых марковских моделях) правдоподобие может быть подсчитанно непосредственно. Главным недостатком этого метода является невозможность его использования в случае графов с циклами, т.к. для циклических графов не существует эффективных алгоритмов подсчета нормировочной константы $Z(X,W)$. Кроме того, последние исследования показывают, что метод максимального правдоподобия, будучи оптимальным в случае известной параметрической модели, проигрывает структурному методу опорных векторов в ситуации, когда истинное распределение не входит в выбранное параметрическое семейство.

Максимизация псевдоправдоподобия. Альтернативой методу максимального правдоподобия является оптимизация весов W с помощью максимизации т.н. псевдоправдоподобия

$$W_{MPL} = \arg\max_W \prod_{i\in V} p(t_i|T_{\setminus i}, X, W),$$

представляющего собой произведение условных вероятностей значения каждой отдельной скрытой переменной при известных значениях всех остальных переменных модели. Легко видеть, что каждый множитель выражается через функцию $L(X,T,W)$ следующим образом

$$p(t_i = k|T_{\setminus i}, X, W) = \frac{\exp\left(\sum_{j\in N(i)}\phi_{ij}(k,t_j) + \phi_i(k)\right)}{\sum_{l=1}^{K}\exp\left(\sum_{j\in N(i)}\phi_{ij}(l,t_j) + \phi_i(l)\right)},$$

где $N(i)$ — множество соседей (вершин, соединенных с данной) i-ой вершины. Во многих случаях такой подход дает приемлемые результаты, но известны ситуации, когда качество обучения оказывалось плохим. Основной проблемой в использовании псевдоправдоподобия является то, что на этапе тестирования условные вероятности для каждой скрытой переменной не будут доступны даже при наличии обученных весов. Таким образом, происходит оптимизация не того функционала, который мог бы гарантировать малую ошибку сегментации при отсутствии информации о значении других скртытых переменных.

Кусочное обучение (piece-wise learning). Одной из простейших альтернатив структурному обучению является кусочное обучение, при котором унарные потенциалы настраиваются отдельно и независимо как в случае стандартной задачи машинного

обучения. Применительно к дискретным переменным это соответствует обучению классификатора на K классов. Аналогично можно провести настройку парных потенциалов, рассматривая каждую пару вершин, соединенных ребром как некоторый объект, описываемый d_2 признаками[10] и относящийся к одному из $K \times K$ классов. Такой способ обучения требует горазхдо меньше времени и часто его оказывается достаточным для получения адекватной вероятностной модели. Также, часто этот способ используется для того, чтобы «нащупать» информативные описания вершин и ребер графа, над которыми затем запускают структурные методы. Основным недостатком метода является невозможность учета взаимосвязей между классами вершин, соединенных ребрами. Заметим, что значения признаков соседних вершин мы можем учесть, дополнив признаковое описание каждой вершины признаками его соседей.

Функциональный градиентный бустинг. Решается задача минимизации величины ξ при ограничениях (9), (10). На каждой итерации находится наиболее нарушаемое ограничение, после Недостаток предыдущего метода – то что он позволяет обучать только линейные функции от параметров. Рассмотрим нелинейное обобщение.

Максимизация произведения маргинальных распределений. При таком подходе веса W ищутся как

$$W_H = \arg\max_W \prod_{i \in V} p(t_i|X, W),$$

где $p(t_i|X, W) = \sum_{T \setminus i} p(T|X, W)$ — маргинальное распределение значений i-ой вершины. Можно показать, что такой подход приводит в минимизации числа неправильно классифицированных вершин графа, при решающем правиле, имеющим вид максимума по маргинальной распределению

$$T = \{t_i\} = \{\arg\max p(t_i|X, W_H)\}.$$

Главной трудностью явялется невозможность подсчета маргинальный распределений для графов с циклами. Даже их приближенная оценка, как правило, сопряжена с большими трудностями чем нахождение наиболее вероятной конфигурации скрытых переменных.

[10]В принципе, никто не запрещает добавить к парным признакам и $2d_1$ признаков вершин, т.к. это информация будет также доступна при подсчете значения парного потенциала.