Семинар 7.
ММП, осень 2012–2013
13 ноября

Темы семинара:

- Линейные методы классификации;
- Метод множителей Лагранжа;
- Оптимальная разделяющая гиперплоскость, SVM.

1 Разбор домашнего задания

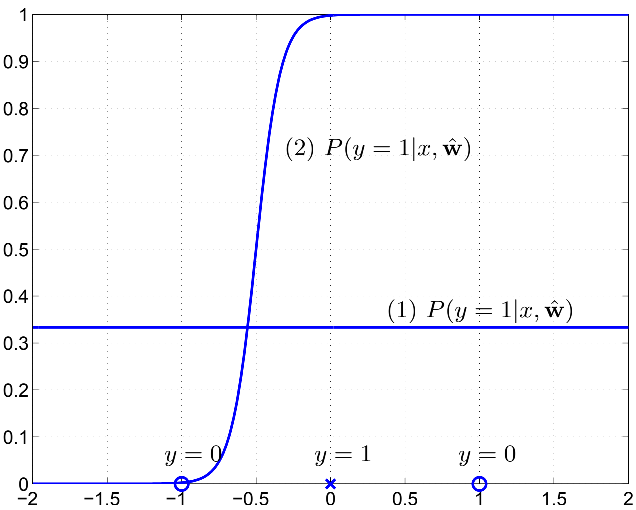

Рис. 1: Обучающие объекты.

Задача. Рассмотрим логистическую регрессию в одномерной задаче $\mathbb{X} = \mathbb{R}$ с двумя классами $\mathbb{Y} = \{0, 1\}$:

$$\mathsf{P}(y = 1|x, \boldsymbol{w}) = \sigma(w_1 x + w_0),$$

где $\sigma(x) = \frac{1}{1+e^{-x}}$ — логистическая сигмоида.

На рисунке 3 приведены две разных функции апостериорных вероятностей $\mathsf{P}(y = 1|x, \boldsymbol{w})$ принадлежности к классу 1, получающихся при различных параметрах $\boldsymbol{w}$.

а) Для каждой из апостериорных вероятностей укажите число ошибок, допускаемых на объектах, приведенных на том же рисунке.
б) Одна из приведенных апостериорных вероятностей соответствует вектору $\boldsymbol{w}$, полученному методом максимизации правдоподобия на объектах, указанных на рисунке. Которая из них?
в) Повлияет ли на решение пункта б) добавление *регуляризатора* $w_1^2/2$ к логарифму функции правдоподобия?

Решение:

а) Как мы знаем, после настройки параметра $\boldsymbol{w}$ логистической регрессии, классификация выглядит следующим образом:

$$a(X) = \arg\max_Y \lambda_Y p(Y|X, \boldsymbol{w}).$$

Мы рассмотрим наиболее частый в задачах классификации случай, когда потери $\lambda_1 = \lambda_0$ совпадают. Как упоминалось на лекциях, к этому случаю ведет наиболее часто используемая в задачах классификации *бинарная* функция потерь $\ell(y_1, y_2) = [y_1 \neq y_2]$. Легко проверить, что модель (1) ошибается на одной точке $X = 0$, а модель (2) — на точке $X = 1$.

б) Напомним, какая оптимизационная задача решается при использовании в логистической регрессии метода максимума правдомодобия:

$$\prod_{i=1}^{\ell} p(y_i|\boldsymbol{x}_i) \to \max_{\boldsymbol{w}}. \tag{1}$$

Все, что нужно сделать — посчитать оптимизируемое выражение для двух моделей (1) и (2). Модель (2) для точки $X = 1$, которая принадлежит классу $Y = 0$, присваивает апостериорную вероятность $p(Y = 1|X = 1) \approx 1$, то есть $p(Y = 0|X = 1) \approx 0$. Таким образом соответствующий этой точке множитель в выражении (1) будет почти равен нулю, и все правдоподобие в результате будет очень близко к нулю. Для модели (1) левая часть выражения (1) равна $\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} > 0$. Значит, ответ — модель (1).

в) Обратим внимание на то, что апостериорная вероятность модели (1) является константой. Вспомним вид апостериорной вероятности в логистической регрессии:

$$p(Y = 1|X) = \sigma(\langle \boldsymbol{w}, X\rangle) = \frac{1}{1 + \exp(-\langle \boldsymbol{w}, X\rangle)}.$$

Поскольку $p(Y = 1|X) = const(X)$, мы приходим к выводу, что для этой модели $w_1 = 0$. А значит регуляризатор не повлияет на решение, поскольку решение задачи без регуляризации уже имеет параметр, минимизирующий предлагаемый регуляризатор.

Замечание. *Зачем в лекциях логарифмическая функция потерь вводится с логарифмом по основанию 2?*

2 Метод множителей Лагранжа (кратко)

Для понимания работы SVM нам понадобится метод множителей Лагранжа — процедура, используемая для поиска условного экстремума функции многих переменных. Здесь мы очень вкратце опишем, в чем состоит этот метод, и приведем лишь результаты, которые нам пригодятся в работе с методом опорных векторов.

2.1 Условие-равенство.

Рассмотрим сначала случай, когда условия принимают вид равенств. Предположим, что мы хотим решить следующую задачу:

$$\begin{cases} f(\boldsymbol{x}) \to \max_{\boldsymbol{x}}; \\ g(\boldsymbol{x}) = 0; \end{cases} \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{1}$$

Условие-уравнение $g(\boldsymbol{x}) = 0$ задает в $\mathbb{R}^n$ $(n-1)$-мерную поверхность, на которой нам и предстоит максимизировать функцию f. Обозначим эту поверхность S_g.

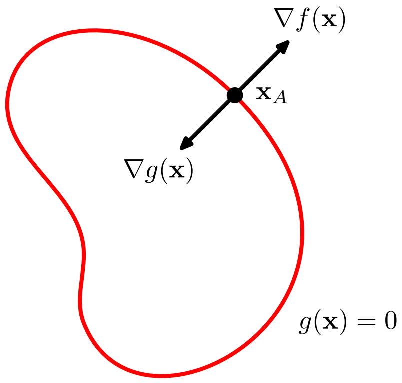

Рис. 2: Поверхность, задающая условие $g(\boldsymbol{x}) = 0$.

Обратим внимание на следующий факт. Рассмотрим точки $\boldsymbol{x}$ и $\boldsymbol{x} + \boldsymbol{\varepsilon}$ из S_g. Разложим функцию g в окрестности точки $\boldsymbol{x}$ в ряд Тейлора:

$$g(\boldsymbol{x} + \boldsymbol{\varepsilon}) = g(\boldsymbol{x}) + \langle \boldsymbol{\varepsilon}, \nabla g(\boldsymbol{x}) \rangle + o(\|\boldsymbol{\varepsilon}\|).$$

При $\|\boldsymbol{\varepsilon}\| \to 0$ это выражение обращается в равенство. Поскольку в этом случае вектор $\boldsymbol{\varepsilon}$ направлен вдоль касательной к поверхности S_g, а также $g(\boldsymbol{x}) = g(\boldsymbol{x} + \boldsymbol{\varepsilon}) = 0$, мы приходим к выводу, что градиент функции g в любой точке поверхности S_g направлен перпендикулярно к ней. Из похожих соображений несложно заключить, что градиент функции f в точке $x^* \in S_g$, в которой она достигает своего максимума, также перпендикулярен поверхности S_g. Заметим, что градиент функции f в искомой точке совершенно не обязан быть нулевым: это было бы справедливо в случае поиска безусловного экстремума. Таким образом в искомой точке градиенты функций f и g параллельны и удовлетворяют равенству

$$\nabla f + \lambda \nabla g = 0 \tag{2}$$

для некоторого числа $\lambda \neq 0$, которое называется *множителем Лагранжа*. Важно заметить, что множитель Лагранжа в данном случае может иметь любой знак.

Введем функцию Лагранжа

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}). \tag{3}$$

Условие (2) теперь можно записать как $\nabla_{\boldsymbol{x}} L = 0$, а условие $g(\boldsymbol{x}) = 0$ — в виде $\frac{\partial L}{\partial \lambda} = 0$. Мы получили *необходимые условия* для решения задачи (1):

> Пусть x^* — решение задачи (1). Тогда найдется число $\lambda^* \neq 0$, удовлетворяющее системе
> $$\begin{cases} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) + \lambda^* \nabla_{\boldsymbol{x}}\, g(\boldsymbol{x}^*) = 0; \\ g(\boldsymbol{x}^*) = 0. \end{cases}$$

Задача. *Найдите решение следующей задачи:*

$$\begin{cases} 1 - x^2 - y^2 \to \max_{x,y}; \\ x + y = 1. \end{cases} \tag{Ex.1}$$

Решение: *выпишем функцию Лагранжа*

$$L(x, y, \lambda) = 1 - x^2 - y^2 + \lambda(x + y - 1).$$

Найдем ее стационарные точки:

$$\begin{cases} -2x + \lambda = 0; \\ -2y + \lambda = 0; \\ x + y = 1. \end{cases}$$

Решение этой системы: $(x = \frac{1}{2}, y = \frac{1}{2}, \lambda = 1)$. *В данном случае это решение исходной задачи (Ex.1).*

2.2 Условие-неравенство.

Что делать, если условия представлены в виде неравенства $g(x) \geqslant 0$, а не равенства?

$$\begin{cases} f(\boldsymbol{x}) \to \max_{\boldsymbol{x}}; \\ g(\boldsymbol{x}) \geqslant 0; \end{cases} \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{4}$$

Возможны два случая.

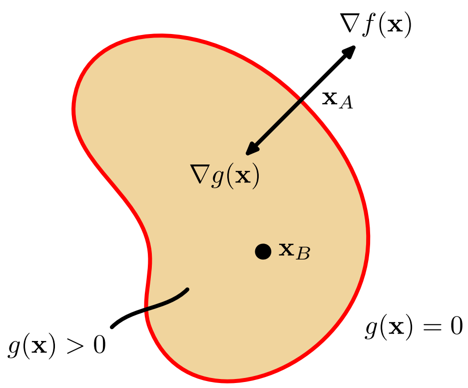

Рис. 3: Условия вида $g(\boldsymbol{x}) \geqslant 0$.

Первый — когда максимум функции f достигается в области $g(\boldsymbol{x}) > 0$. В этом случае условие $g(x) \geqslant 0$ не играет роли и называется *неактивным*. Решение задачи (1) совпадает с решением задачи поиска безусловного экстремума и записывается в виде $\nabla f(\boldsymbol{x}) = \mathbf{0}$. Это решение выражается стационарной точкой функции Лагранжа (3) для $\lambda = 0$.

Во втором случае условный экстремум функции f лежит на поверхности $g(\boldsymbol{x}) = 0$ (которую мы снова будем обозначать S_g). Условие $g(x) \geqslant 0$ в этом случае называется *активным*. В этом случае решение записывается так же, как в задаче с условием-равенством — в виде стационарной точки функции Лагранжа для $\lambda \neq 0$. В отличие от задачи с условием-равенством на этот раз знак множителя Лагранжа важен: точка на S_g будет решением задачи только в том случае, когда градиенты функций f и g противоположны в ней (иначе решение будет достигаться где-то в области $g(\boldsymbol{x}) > 0$). Таким образом выполнено $\nabla f(\boldsymbol{x}) = -\lambda \nabla g(\boldsymbol{x})$ для некоторого $\lambda > 0$.

Мы грубо обрисовали, откуда берутся следующие необходимые условия оптимальности решения задачи (4).

> Предположим, что в задаче (4) f и g вогнуты. Пусть $\boldsymbol{x}^*$ — решение задачи (4). Тогда найдется число λ^*, для которого выполнено
>
> $$\begin{cases} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*) + \lambda^* \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^*) = 0; \\ \lambda^* \geqslant 0; \\ g(\boldsymbol{x}^*) \geqslant 0; \\ \lambda^* g(\boldsymbol{x}^*) = 0. \text{ т.н. } \textit{условие дополняющей нежесткости.} \end{cases} \tag{5}$$

Отметим, что если мы хотим минимизировать функцию f, а не максимизировать, её градиент во втором из рассмотренных случаев будет направлен в сторону области $g(\boldsymbol{x}) > 0$, то есть сонаправлен с градиентом функции g. В этом случае мы определяем функцию Лагранжа $L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$. Остальные условия в (5) при этом остаются без изменений.

Приведем условия, при которых система (5) становится необходимыми и достаточными условиями оптимального решения задачи (4).

Теорема 1. *Пусть мы решаем задачу условной максимизации функции f, а условия записываются в виде неравенств $g_i(\boldsymbol{x}) \geqslant 0$, $i = 1, \ldots, m$. Тогда условия (5) становятся необходимыми и достаточными для решения этой задачи, если внутренность множества ограничений не пуста, т. е. найдется $\boldsymbol{x}_0$: $g(\boldsymbol{x}_0) > 0$ (так называемые условия Слейтера).*

Задача. *Решите следующую задачу условной оптимизации:*

$$\begin{cases} -(x-4)^2 - (y-4)^2 \to \max_{x,y}; \\ x + y \leqslant 4; \\ x + 3y \leqslant 9. \end{cases}$$

Решение. *Выпишем функцию Лагранжа:*

$$L(x, y, \lambda_1, \lambda_2) = -(x-4)^2 - (y-4)^2 + \lambda_1(4 - x - y) + \lambda_2(9 - x - 3y).$$

Условия Куна–Таккера запишутся в виде:

$$\begin{cases} -2(x-4)-\lambda_1-\lambda_2=0; \\ -2(y-4)-\lambda_1-3\lambda_2=0; \\ x+y\leqslant 4,\ \lambda_1\geqslant 0,\ \lambda_1(x+y-4)=0; \\ x+3y\leqslant 9,\ \lambda_2\geqslant 0,\ \lambda_2(x+3y-9)=0. \end{cases}$$

Решая их, рассмотрим 4 случая:

- $x+y=4,\ \ x+3y=9,\ \ \lambda_1>0,\ \ \lambda_2>0.$
  Два эти уравнения дают $(x=\frac{3}{2}, y=\frac{5}{2})$. *После подстановки в первые два уравнения условий Куна–Таккера, получаем*

  $$\begin{cases} -2(\frac{3}{2}-4)-\lambda_1-\lambda_2=0; \\ -2(\frac{5}{2}-4)-\lambda_1-3\lambda_2=0, \end{cases}$$

  откуда $\lambda_2=-1$, *что противоречит принятым условиям.*

- $x+y=4,\ \ x+3y<9,\ \ \lambda_1>0,\ \ \lambda_2=0.$
  Подстановка $\lambda_2=0$ *в первые два уравнения условий Куна–Таккера вместе с уравнением* $x+y=4$ *дают решение* $(x=2, y=2, \lambda_1=4, \lambda_2=0)$. *Эти решения удовлетворяют всем условиям Куна–Таккера.*

- *Проверьте, что оставшихся два случая ведут к противоречиям, так же как первый случай.*

Поскольку условия теоремы 1 выполнены, найденная точка является решением исходной задачи.

2.3 Смешанный набор условий.

Полученные необходимые условия для случаев ограничений-равенств и ограничений-неравенств понятным образом обобщаются на задачи оптимизации с произвольным конечным числом смешанных условий. Здесь мы рассмотрим условную минимизацию функции, которая нам пригодится в следующих рассуждениях:

$$\begin{cases} f(\boldsymbol{x})\to\min_{\boldsymbol{x}}; \\ g_i(\boldsymbol{x})=0; \quad i=1,\dots,p\,; \qquad\qquad \boldsymbol{x}\in\mathbb{R}^n, \\ h_j(\boldsymbol{x})\leqslant 0; \quad j=1,\dots,q, \end{cases}$$

где функции f, g_i, h_i выпуклы. Отметим важный момент: в этом случае любой локальный оптимум задачи является также ее глобальным оптимумом.

Тогда по теореме Куна–Таккера, если $\boldsymbol{x}^*$ — решение этой задачи, то найдутся числа $(\lambda_1^*,\dots,\lambda_p^*,\mu_1^*,\dots,\mu_q^*)$, для которых выполнена следующая система:

$$\begin{cases} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*)+\sum_{i=1}^{p}\lambda_i^*\nabla_{\boldsymbol{x}}\, g_i(\boldsymbol{x}^*)+\sum_{j=1}^{q}\mu_j^*\nabla_{\boldsymbol{x}} h_j(\boldsymbol{x}^*)=0; \\ g_i(\boldsymbol{x}^*)=0, \quad i=1,\dots,p\,; \\ h_j(\boldsymbol{x}^*)\geqslant 0, \quad j=1,\dots,q; \\ \mu_j^*\geqslant 0, \quad j=1,\dots,q; \\ \mu_j^* h_j(\boldsymbol{x}^*=0, \quad j=1,\dots,q. \end{cases}$$

В случае, если h_j линейны и внутренность множества ограничений не пуста, то эти условия также становятся достаточными условиями оптимального решения.

3 SVM: случай линейной разделимости.

Мы будем рассматривать задачу классификации с двумя классами $\mathbb{Y} = \{-1, +1\}$ в $\mathbb{R}^n$. Рассмотрим сначала случай, когда точки двух классов из обучающей выборки X^ℓ линейно разделимы. В этом случае рассматриваемый метод называется *методом оптимальной гиперплоскости* и состоит, как вы знаете, в поиске линейного классификатора

$$a(\boldsymbol{x}) = \operatorname{sgn}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle - w_0), \tag{3.1}$$

безошибочно классифицирующего обучающую выборку. Причем из всех гиперплоскостей $(\boldsymbol{w}, w_o)$, не допускающих ошибку на обучении, метод выбирает ту, для которой расстояние до ближайшей точки обучающей выборки максимлаьно:

$$\min_{i=1,\dots,\ell} \frac{y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0)}{\|\boldsymbol{w}\|} \to \max_{\boldsymbol{w},\, w_0}.$$

Задача. *Докажите, что записанная задача соответствует максимизации зазора — расстояния от гиперплоскости до ближайшей к ней точки обучающей выборки.*

Одновременное умножение вектора нормали $\boldsymbol{w}$ и порога w_0 на одно и то же число не меняет расстояний от точек до разделяющей гиперплоскости. Воспользовавшись этим свойством, мы полагаем отступ $M_i = y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0)$ точки, ближайшей к гиперплоскости, равным 1. Таким образом мы получаем задачу:

$$\begin{cases} \frac{1}{2}\|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w},\, w_0}; \\ y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0) \geqslant 1, \quad i = 1, \dots, \ell. \end{cases} \tag{3.2}$$

Задача. *Докажите, что в задаче (3.2) всегда будет по крайней мере одно активное ограничение.*

Решение: *очевидно, поскольку всегда есть точка, ближайшая к гиперплоскости.*

Задача. *Допустим, мы решили задачу (3.2). Что можно сказать о числе активных ограничений в этом случае?*

Решение. *В этом случае их будет не меньше 2-х. Что тоже очевидно.*

Задача. *Решите задачу (3.2).*

Видим, что для задачи (3.2) мы можем применить рассмотренные в прошлом разделе методы для поиска оптимального решения. Функция Лагранжа для задачи (3.2) выглядит следующим образом:

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{\ell} \lambda_i \big(y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0) - 1 \big),$$

где λ_i — множители Лагранжа. Условия Куна-Такера примут вид:

$$\begin{cases} \boldsymbol{w} = \sum_{i=1}^{\ell} \lambda_i y_i \boldsymbol{x}_i; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0; \\ \lambda_i \geqslant 0, \quad i = 1, \ldots, \ell; \\ y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0) \geqslant 1, \quad i = 1, \ldots, \ell; \\ \lambda_i \big(y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - w_0) - 1 \big) = 0, \quad i = 1, \ldots, \ell. \end{cases} \tag{3.3}$$

Задача. *Выпишите вид итогового классификатора, получаемого решением задачи (3.2), используя множители Лагранжа λ_i.*

Первое из условий (3.3) дает

$$a(\boldsymbol{x}) = \operatorname{sgn} \left\{ \sum_{i=1}^{\ell} \lambda_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle - w_0 \right\}.$$

Задача. *К чему ведут условия дополняющей нежесткости в данном случае? На какие множества разбиваются точки обучающей выборки?*

Задача. *Почему метод оптимальной разделяющей гиперплоскости называют разреженным по объектам?*

Задача. *Как получить значение порога w_0?*