Семинар 9.
ММП, осень 2012–2013
27 ноября

Илья Толстихин
iliya.tolstikhin@gmail.com

Темы семинара:

- Линейные методы классификации;
- SVM;
- Ядровая функция, ядровой переход.

1 Решения домашнего задания

Задача. *Покажите, что для задачи условной оптимизации soft-margin SVM можно записать эквивалентную задачу безусловной оптимизации, воспользовавшись условиями* $\xi_i \geqslant \max\big(0, 1 - y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - w_0)\big)$*, которые выполнены для всех* $i = 1, \ldots, \ell$*. Какую связь этой задачи безусловной оптимизации с минимизацией эмпирического риска вы видите? Какая функция потерь используется? Как эта задача связана с решением логистической регрессии?*

Задача. *Докажите, что половина ширины разделяющей полосы* ρ*, соответствующей оптимальной разделяющей гиперплоскости для случая hard-margin, вычисляется следующим образом:*

$$\frac{1}{\rho^2} = \sum_{i=1}^{\ell} \lambda_i,$$

где λ_i *— решения следующей задачи, двойственной к задаче условной оптимизации hard-margin SVM:*

$$\begin{cases} \sum_{i=1}^{\ell} \lambda_i - \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \to \max_{\boldsymbol{\lambda}}; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0; \\ \lambda_i \geqslant 0, \quad i = 1, \ldots, \ell. \end{cases}$$

2 Ядровой переход

На прошлых семинарах мы подробно изучили свойства оптимальной разделяющей гиперплоскости (hard-margin SVM) и общего случая метода опорных векторов

при линейно неразделимой выборки (soft-margin SVM) в задаче классификации с двумя классами и пространством объектов $\mathbb{X} = \mathbb{R}^n$. Мы выписали двойственные задачи для каждого из случаев:

- Оптимальная разделяющая гиперплоскость:

$$\begin{cases} \sum_{i=1}^{\ell} \lambda_i - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \to \max_{\boldsymbol{\lambda}}; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0; \\ \lambda_i \geqslant 0, \ i = 1, \ldots, \ell. \end{cases} \tag{1}$$

- Случай линейно неразделимой выборки:

$$\begin{cases} \sum_{i=1}^{\ell} \lambda_i - \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \to \max_{\boldsymbol{\lambda}}; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0; \\ 0 \leqslant \lambda_i \leqslant C, \ i = 1, \ldots, \ell. \end{cases} \tag{2}$$

Также мы показали, что разделяющая поверхность настроенного классификатора $\langle \boldsymbol{w}^*, \boldsymbol{x} \rangle = w_0^*$ в обоих случаях выражается в виде

$$\sum_{i=1}^{\ell} \lambda_i^* y_i \langle \boldsymbol{x}_i, x \rangle = w_0^*. \tag{3}$$

Таким образом, при использовании метода опорных векторов как решаемая задача оптимизации, так и результирующее решающее правило зависят лишь от скалярных произведений объектов обучающей выборки и новых классифицируемых объектов, но не от самих признаков.

Рассмотрим некоторое отображение $\psi \colon \mathbb{X} \to \mathcal{H}$ исходного пространства объектов в новое пространство $\mathcal{H}$. Если мы знаем, как вычислять скалярное произведение $\langle \boldsymbol{x}', \boldsymbol{x}'' \rangle_{\mathcal{H}} = \langle \psi(\boldsymbol{x}'), \psi(\boldsymbol{x}'') \rangle$ в новом пространстве $\mathcal{H}$, мы сразу же получаем возможность строить линейные классификаторы в новом пространстве $\mathcal{H}$, пользуясь изученными на последних двух семинарах методами. При этом при соответствующем выборе отображения ψ настроенная линейная разделяющая поверхность в новых координатах $\mathcal{H}$ будет соответствовать *нелинейной* разделяющей поверхности в старых. Пространство $\mathcal{H}$ Мы будем называть *спрямляющим*.

Предположим, что скалярное произведение в выбранном новом пространстве $\mathcal{H}$ выражается в виде

$$\langle \boldsymbol{x}', \boldsymbol{x}'' \rangle_{\mathcal{H}} = \langle \psi(\boldsymbol{x}'), \psi(\boldsymbol{x}'') \rangle = K_{\mathcal{H}}(\boldsymbol{x}', \boldsymbol{x}''), \quad K_{\mathcal{H}} \colon \mathbb{X}^2 \to \mathbb{R}. \tag{4}$$

Функцию $K_{\mathcal{H}}(\cdot, \cdot)$ мы будем называть *ядровой*. Получается, что для обучения методом опорных векторов необязательно в явном виде указывать спрямляющее пространство $\mathcal{H}$: достаточно лишь задать ядровую функцию K. Конечно, не любая функция будет ядровой (не для любой K найдется отображение $\psi \colon \mathbb{X} \to \mathcal{H}$, такое что (4) выполнено). На лекции были введены *условия Мерсера* — необходимые и достаточные условия, при которых K является ядровой функцией:

1. симметричность — $K(\boldsymbol{x}', \boldsymbol{x}'') = K(\boldsymbol{x}'', \boldsymbol{x}')$;

2. положительная определенность — для всех g: $\int g^2(\boldsymbol{x})d\boldsymbol{x} < \infty$ выполнено

$$\int K(\boldsymbol{x}', \boldsymbol{x}'')g(\boldsymbol{x}')g(\boldsymbol{x}'')d\boldsymbol{x}'d\boldsymbol{x}'' \geqslant 0.$$

Задача. *Рассмотрим $\mathbb{X} = \mathbb{R}^2$. Докажите, что $K(\boldsymbol{x}', \boldsymbol{x}'') = \langle \boldsymbol{x}', \boldsymbol{x}'' \rangle^2$ — функция ядра. Найдите размерность соответствующего этой ядровой функции спрямляющего пространства и отображение ψ.*
Решение:

$$K(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{u}, \boldsymbol{v} \rangle^2 = (\boldsymbol{v}_1\boldsymbol{u}_1 + \boldsymbol{v}_2\boldsymbol{u}_2)^2 = \boldsymbol{v}_1^2\boldsymbol{u}_1^2 + \boldsymbol{v}_2^2\boldsymbol{u}_2^2 + 2\boldsymbol{v}_1\boldsymbol{u}_1\boldsymbol{v}_2\boldsymbol{u}_2 =$$
$$= \langle (\boldsymbol{u}_1^2, \boldsymbol{u}_2^2, \sqrt{2}\boldsymbol{u}_1\boldsymbol{u}_2), (\boldsymbol{v}_1^2, \boldsymbol{v}_2^2, \sqrt{2}\boldsymbol{v}_1\boldsymbol{v}_2) \rangle.$$

Применение спрямляющего пространства $\psi(x, y) = (x, y, x^2 + y^2)$ для отделения внутренней части эллипса от его внешней части проиллюстрировано в довольно примитивном, но наглядном видео [4].

Одной из часто используемых на практике ядровых функций является *гауссовская радиальная базисная функция*:

$$K(\boldsymbol{x}', \boldsymbol{x}'') = \exp\left(-\frac{\|\boldsymbol{x}' - \boldsymbol{x}''\|^2}{2\sigma^2}\right), \tag{5}$$

где $\sigma > 0$.

Задача. *Пользуясь конструктивными способами построения ядер, приведенными на лекции (раздел 4.5.3), докажите, что (5) действительно является ядровой функцией.*

Рассмотрим ряд интересных свойств этого ядра. Напомним следующий факт из курса линейной алгебры:

Теорема 2.1 *Рассмотрим матрицу Грамма системы векторов $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ евклидова пространства:*

$$G(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) = \begin{bmatrix} \langle \boldsymbol{a}_1, \boldsymbol{a}_1 \rangle & \langle \boldsymbol{a}_1, \boldsymbol{a}_2 \rangle & \ldots & \langle \boldsymbol{a}_1, \boldsymbol{a}_k \rangle \\ \ldots & \ldots & \ldots & \ldots \\ \langle \boldsymbol{a}_k, \boldsymbol{a}_1 \rangle & \langle \boldsymbol{a}_k, \boldsymbol{a}_2 \rangle & \ldots & \langle \boldsymbol{a}_k, \boldsymbol{a}_k \rangle \end{bmatrix}.$$

Система векторов $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ линейно независима тогда и только тогда, когда матрица Грамма невырождена $\det G(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \neq 0$.

Следующая теорема утверждает, что матрица грамма, соответствующая гауссовскому ядру (5), для k различных векторов имеет полный ранг:

Теорема 2.2 *Пусть $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell$ — различные точки пространства объектов $\mathbb{X} = \mathbb{R}^n$. Тогда матрица Грамма*

$$G_{i,j} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right), \quad i, j = 1, \ldots, \ell,$$

где $\sigma > 0$, невырождена.

Задача. *Желающие могут попробовать доказать этот факт.*

Задача. *Докажите, что гауссовское ядро соответствует спрямляющему пространству бесконечномерной размерности.*

С учетом теорем (2.1) и (2.2) можно заключить, что любые k различных точек обучающей выборки в спрямляющем пространстве $\mathcal{H}$, соответствующем гауссовскому ядру (5), образуют базис его k-мерного подпространства. Отсюда следует, что если не ограничивать число различных точек в обучающей выборке, то гауссовское ядро соответствует спрямляющему пространству бесконечной размерности.

Задача. *Докажите, что метод оптимальной разделяющей гиперплоскости, использующий гауссовское ядро, может безошибочно классифицировать произвольную обучающую выборку, состоящую из ℓ различных объектов и двух классов.*

Зафиксируем произвольный параметр $\sigma > 0$ и обозначим соответствующее гауссовское ядро $K_G(\boldsymbol{x}', \boldsymbol{x}'')$. Покажем существование вектора $\boldsymbol{w} \in \mathcal{H}$ в спрямляющем пространстве (соответствующем зафиксированному гауссовскому ядру) для которого выполнена следующая система:

$$\langle \boldsymbol{w}, \psi(\boldsymbol{x}_i) \rangle = y_i, \quad i = 1, \ldots, \ell, \tag{6}$$

где ψ — отображение в спрямляющее пространство, соответствующее гауссовскому ядру. Очевидно, что отсюда будет следовать, что обучающая выборка в соответствующем спрямляющем пространстве линейно разделима, что и требовалось доказать.

Мы уже установили, что вектора $\psi(\boldsymbol{x}_1), \ldots, \psi(\boldsymbol{x}_\ell)$ образуют в спрямляющем пространстве линейно независимую систему. Линейная оболочка этих векторов — ℓ-мерное подпространство спрямляющего пространства $\mathcal{H}$. Будем искать указанный вектор $\boldsymbol{w}$ в этой ℓ-мерной линейной оболочке, то есть положим

$$\boldsymbol{w} = \sum_{i=1}^{\ell} a_i \psi(\boldsymbol{x}_i).$$

Тогда условия (6) запишутся в виде

$$\sum_{j=1}^{\ell} a_j \langle \psi(\boldsymbol{x}_j), \psi(\boldsymbol{x}_i) \rangle = y_i, \quad i = 1, \ldots, \ell.$$

Используя гауссовское ядро перепишем условия в виде

$$\sum_{j=1}^{\ell} a_j K_G(\boldsymbol{x}_j, \boldsymbol{x}_i) = y_i, \quad i = 1, \ldots, \ell,$$

или в матричном виде

$$\begin{pmatrix} K_G(\boldsymbol{x}_1, \boldsymbol{x}_1) & K_G(\boldsymbol{x}_1, \boldsymbol{x}_2) & \ldots & K_G(\boldsymbol{x}_1, \boldsymbol{x}_k) \\ \ldots & \ldots & \ldots & \ldots \\ K_G(\boldsymbol{x}_k, \boldsymbol{x}_1) & K_G(\boldsymbol{x}_k, \boldsymbol{x}_2) & \ldots & K_G(\boldsymbol{x}_k, \boldsymbol{x}_k) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_\ell \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}.$$

Из теоремы (2.2) следует, что данная СЛАУ имеет решение, поскольку матрица системы невырождена. Откуда мы получаем существование искомого вектора $\boldsymbol{w}$.

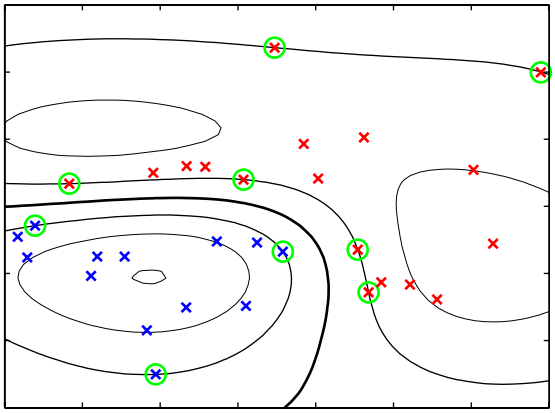
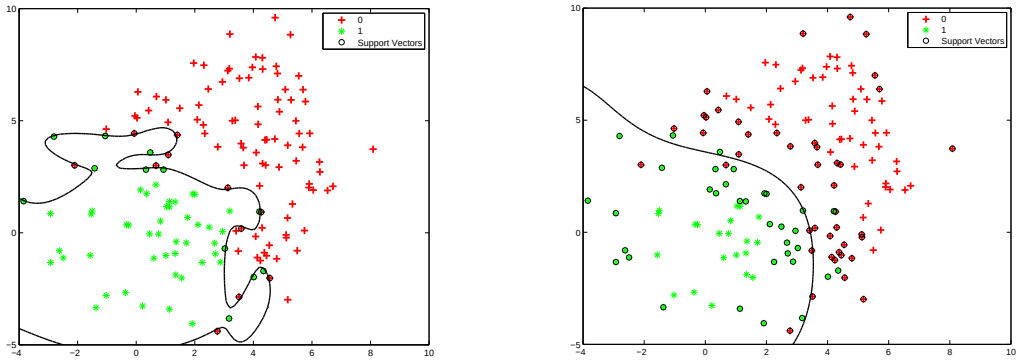


Рис. 1: Оптимальная разделяющая гиперплоскость с гауссовским ядром.

На рисунке 1 (из книги C. Bishop) приведена модельная обучающая выборка с двумя классами. Толстой черной линией изображена разделяющая поверхность, построенная методом опорных векторов с гауссовской ядровой функцией. Зеленым обведены опорные векторы. Также на рисунке приведены линии уровня функции $y(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle - w_0$, полученной в результате обучения. В том числе — линии единичных отступов $y(\boldsymbol{x}) = 1$ и $y(\boldsymbol{x}) = -1$.

Итак, описанный метод строит классификаторы, не допускающие ошибок на обучающей выборке. Однако при решении прикладных задач в обучающей выборке часто присутствуют шумы и погрешности измерений признаков. Это означает, что мы не всегда хотим строить классификатор, идеально разделяющий объекты обучающей выборки. Оптимальная разделяющая гиперплоскость с гауссовской ядровой функцией в исходном признаковом пространстве будет иметь очень извилистый контур, как показано на рисунке 2.а. На практике гауссовская ядровая функция используется в SVM с регуляризацией (soft-margin) с последующим подбором параметра регуляризации C (фигурирующего в задаче (2)) и *ширины* гауссовского ядра σ^2 (в определении 5) с помощью, например, скользящего контроля. На рисунке 2.б приведена та же выборка с разделяющей поверхностью, полученной решением soft-margin задачи (2) при $C = 1$ и $\sigma = 1$.



а) б)

Рис. 2: Разделяющие поверхности метода опорных векторов.

В области методов оптимизации для SVM существуют результаты, позволяющие проводить обучение с линейным ядром ($K(\boldsymbol{x}', \boldsymbol{x}'') = \langle \boldsymbol{x}', \boldsymbol{x}'' \rangle$) гораздо быстрее, чем с каким-либо другим ядром. Поэтому часто (например, при больших обучающих выборках) ручной подбор явного вида спрямляющего пространств, соответствующего отображения $\psi(\cdot)$ и нового признакового описания является важной задачей.

Известно, что метод опорных векторов чувствителен к нормировке признаков объектов. Часто нормировка ведет к существенному улучшению точности классификатора. Нормировку можно проводить в исходном пространстве признаков: например, *стандартизировать* все признаки — вычитать из столбца признака среднее выборочное его значений и делить столбец на корень из выборочной дисперсии. Другой вариант — нормировать сами объекты: в исходном признаковом пространстве $\tilde{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|$, либо в спрямляющем пространстве, которому соответствует выбранная ядровая функция $K(\cdot, \cdot)$.

Ядра для строк и последовательностей. Рассмотрим в завершении еще один пример специфической ядровой функции, предназначенной для работы со строками и последовательностями конечной длины. Допустим, нам дан фиксированный алфавит $\mathcal{D}$ размера $|\mathcal{D}|$, и мы хотим решать задача, связанные с последовательностями конечной длины, состоящими из букв этого алфавита. Например — биологические задачи, связанные *с прогнозированием вторичной и третичной структуры белка*. Эти задачи заключаются в прогнозировании *вторичной* и *третичной* структуры белка по его *первичной* структуре — последовательности аминокислот, составляющих его. Это короткое видео [5] очень наглядно изображает все уровни структуры белка.

Приведем несколько ядровых функций, часто используемых на практике для решения таких задач. Следующая ядровая функция называется *спектральным ядром (spectrum kernel)* [6] и учитывает количество подстрок длины n, одновременно встречающихся в обоих строках $\boldsymbol{s}_1$ и $\boldsymbol{s}_2$:

$$K_n^{spec}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \langle \varphi_n(\boldsymbol{s}_1), \varphi_n(\boldsymbol{s}_2) \rangle,$$

где φ_n — отображение строк в пространство размерности $|\mathcal{D}|^n$. Каждая размерность в нем соответствует одной из $|\mathcal{D}|^n$ различных последовательностей букв длины n и равна числу раз, которое эта подстрока встречается в исходной строке.

Задача. *Указанное ядро связано с экспоненциальным по n числом признаков в спрямляющем пространстве. Убедитесь, что тем не менее для его использования совершенно не обязательно вычислять все $|\mathcal{D}|^n$ значений признаков и достаточно линейного по длине исходных строк времени.*

Как одновременно учитывать подпоследовательности разной длины? С этим справляется следующее *ядро смешанных спектров (mixed spectrum kernel)*

$$K_d^{mixspec}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sum_{i=1}^{d} \beta_i K_i^{spec}(\boldsymbol{s}_1, \boldsymbol{s}_2),$$

где β_i — неотрицательные веса соответствующих длин подстрок.

Наконец, в ряде случаев бывает важным также учитывать расположение подстрок в исходных строках. Например, известно, что те или иные элементы вторичных структур белка часто встречаются рядом с определенными последовательностями

аминокислот. Следующая модификация рассмотренных спектральных ядер справляется с этой задачей для строк $\boldsymbol{s}_1$ и $\boldsymbol{s}_2$ одинаковой длины L:

$$K_d^{WD}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sum_{j=1}^{L} \sum_{i=1}^{d} \beta_i K_i^{spec}(\boldsymbol{s}_1^{[j:j+i]}, \boldsymbol{s}_2^{[j:j+i]}),$$

где $\boldsymbol{s}^{[i:j]}$ — подстрока строки $\boldsymbol{s}$ длиной $j - i + 1$, начиная с i-й позиции.

Список литературы

[1] *B. A. Scholkopf, A. J. Smola.* Learning with kernels. — MIT Press, 2002.

[2] *J. Shawe-Taylor, N. Cristianini* Kernel methods for pattern analysis. — Cambridge University Press, 2004.

[3] *C. Bishop.* Pattern Recognition and Machine Learning. — Springer, 2006.

[4] http://www.youtube.com/watch?v=3liCbRZPrZA.

[5] http://www.youtube.com/watch?v=lijQ3a8yUYQ.

[6] *Asa Ben-Hur, Cheng Soon Ong et. al.* Tutorial: Support Vector Machines and Kernels for Computational Biology, 2008.
http://svmcompbio.tuebingen.mpg.de.