Multiclass Kernel-based Vector Machines

Неформальная постановка задачи

Присвоить объектам метки из конечного множества.

Множество бинарных классификаторов

Идея:

- Построить множество бинарных классификаторов, каждый из которых отделяет один класс от других.

Плюсы:

- Простая и мощная структура.

Минусы:

- Не учитывается корреляция между различными классами.

Формализация задачи

- Обучающая выборка $S = \{(\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m)\}$
- $\bar{x}_i$ содержится в $\mathcal{X} \subseteq \Re^n$, y_i принадлежит $\mathcal{Y} = \{1, \ldots, k\}$
- Нужно построить классификатор $H : \mathcal{X} \to \mathcal{Y}$

Классификатор

- Будем рассматривать классификатор вида

$$H_{\mathbf{M}}(\bar{x}) = \arg\max_{r=1}^{k}\{\bar{M}_r \cdot \bar{x}\}$$

- $\mathbf{M}$ матрица размера $k \times n$, $\bar{M}_r$ строка с номером *r* в матрице $\mathbf{M}$

Бинарный случай

Линейный классификатор определяет объект $\bar{x}$ к классу 1, если $\bar{w} \cdot \bar{x} > 0$, и к классу 2 иначе.

В данном случае матрица, используемая в классификаторе, размера $2 \times n$, где $\bar{M}_1 = \bar{w}$ и $\bar{M}_2 = -\bar{w}$.

Функционал ошибки

$$\epsilon_S(M) = \frac{1}{m}\sum_{i=1}^{m}[H_{\mathrm{M}}(x_i) \neq y_i]$$

- Дискретный - плохо

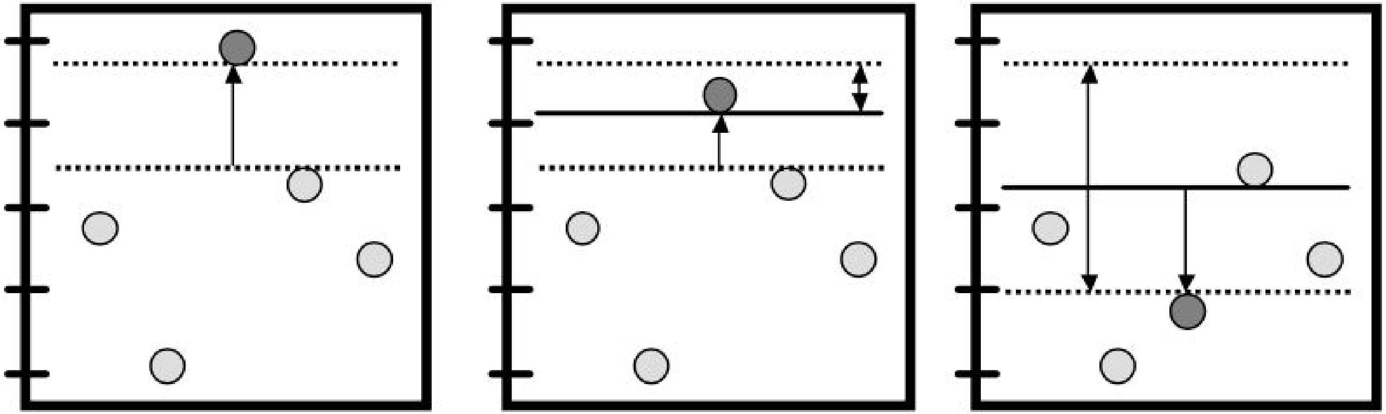
Классификатор с использованием ядра

- Заменим $[\![H_{\mathrm{M}}(x) \neq y]\!]$ на

$$\max_r \{\bar{M}_r \cdot \bar{x} + 1 - \delta_{y,r}\} - \bar{M}_y \cdot \bar{x}$$

- $\delta_{p,q}$ равно 1, если $p = q$ и 0 иначе.

Пример



$$\max_r \{\bar{M}_r \cdot \bar{x} + 1 - \delta_{y,r}\} - \bar{M}_y \cdot \bar{x}$$

Верхняя оценка функционала ошибки

$$\epsilon_S(M) \leq \frac{1}{m} \sum_{i=1}^{m} \left[\max_r \{ \bar{M}_r \cdot \bar{x}_i + 1 - \delta_{y_i,r} \} - \bar{M}_{y_i} \cdot \bar{x}_i \right]$$

Линейно разделимый случай

Будем говорить, что множество разделимо, если существует матрица $\mathbf{M}$, такая что

$$\forall i \quad \max_r\{\bar{M}_r \cdot \bar{x}_i + 1 - \delta_{y_i,r}\} - \bar{M}_{y_i} \cdot \bar{x}_i = 0$$

Эквивалентная запись

$$\forall i, r \quad \bar{M}_{y_i} \cdot \bar{x}_i + \delta_{y_i,r} - \bar{M}_r \cdot \bar{x}_i \geq 1$$

Регуляризатор

- Определим l_2-норму матрицы $\mathbf{M}$, $\|M\|_2^2 = \|(\bar{M}_1, \ldots, \bar{M}_k)\|_2^2 = \sum_{i,j} M_{i,j}^2$
- Получаем следующую оптимизационную задачу

$$\begin{aligned} &\min_{M} \quad \frac{1}{2}\|M\|_2^2 \\ &\text{subject to}: \quad \forall i, r \quad \bar{M}_{y_i} \cdot \bar{x}_i + \delta_{y_i,r} - \bar{M}_r \cdot \bar{x}_i \geq 1 \end{aligned}$$

Линейно неразделимый случай

- Добавим штраф $\xi_i \geq 0$

$$\forall i \quad \max_r \{\bar{M}_r \cdot \bar{x}_i + 1 - \delta_{y_i,r}\} - \bar{M}_{y_i} \cdot \bar{x}_i = \xi_i$$

Задача оптимизации

$$\min_{M,\xi} \quad \frac{1}{2}\beta\|M\|_2^2 + \sum_{i=1}^{m}\xi_i$$

$$\text{subject to :} \quad \forall i, r \quad \bar{M}_{y_i} \cdot \bar{x}_i + \delta_{y_i,r} - \bar{M}_r \cdot \bar{x}_i \geq 1 - \xi_i$$

- $\beta > 0$ регуляризационная константа

Решение задачи

- Лагранжиан оптимизационной задачи

$$
\begin{aligned}
\mathcal{L}(M, \xi, \eta) \quad &= \quad \frac{1}{2}\beta \sum_{r} \|\bar{M}_r\|_2^2 + \sum_{i=1}^{m} \xi_i \\
&\quad + \sum_{i,r} \eta_{i,r} \left[\bar{M}_r \cdot \bar{x}_i - \bar{M}_{y_i} \cdot \bar{x}_i - \delta_{y_i,r} + 1 - \xi_i\right] \\
\text{subject to :} \quad & \quad \forall i, r \quad \eta_{i,r} \geq 0 \quad .
\end{aligned}
$$

Решение задачи

- $$\frac{\partial}{\partial \xi_i}\mathcal{L} = 1 - \sum_r \eta_{i,r} = 0 \qquad \Rightarrow \qquad \sum_r \eta_{i,r} = 1$$

- $$\begin{aligned} \frac{\partial}{\partial \bar{M}_r}\mathcal{L} &= \sum_i \eta_{i,r}\bar{x}_i - \sum_{i,y_i=r} \underbrace{\left(\sum_q \eta_{i,q}\right)}_{=1} \bar{x}_i + \beta \bar{M}_r \\ &= \sum_i \eta_{i,r}\bar{x}_i - \sum_i \delta_{y_i r}\bar{x}_i + \beta \bar{M}_r = 0 \quad , \end{aligned}$$

- $$\bar{M}_r = \beta^{-1}\left[\sum_i (\delta_{y_i,r} - \eta_{i,r})\bar{x}_i\right]$$

Решение задачи

- Каждая строка $\bar{M}_r$ в матрице $\mathbf{M}$ есть линейная комбинация $\bar{x}_i$ с $\delta_{y_i,r} - \eta_{i,r}$ весами
- $\bar{x}_i$ будем называть опорным паттерном, если есть строка r с ненулевым коэффициентом
- Каждому $\bar{x}_i$ сопоставляется множество $\{\eta_{i,1}, \eta_{i,2}, \ldots, \eta_{i,k}\}$ причем $\eta_{i,1}, \ldots, \eta_{i,k} \geq 0$ и $\sum_r \eta_{i,r} = 1$
- Каждое множество будем рассматривать как вероятностное распределение над метками $\{1 \ldots k\}$
- $\bar{x}_i$ - опорный паттерн, если соответствующее распределение не сконцентрировано над правильным ответом

Решение задачи

- Преобразуем Лагранжиан, оставив только двойственные переменные

$$\mathcal{Q}(\eta) = -\frac{1}{2}\beta^{-1}\sum_{i,j}(\bar{x}_i \cdot \bar{x}_j)\left[\sum_r(\delta_{y_i,r} - \eta_{i,r})(\delta_{y_j,r} - \eta_{j,r})\right] - \sum_{i,r}\eta_{i,r}\delta_{y_i,r}$$

Решение задачи

- Обозначим за $\bar{1}_i$ вектор с единицей на i месте и с нулями на других местах
- Обозначим за $\bar{1}$ единичный вектор
- Тогда наша задача будет выглядеть так

$$\max_{\eta} \qquad \mathcal{Q}(\eta) = -\frac{1}{2}\beta^{-1}\sum_{i,j}(\bar{x}_i \cdot \bar{x}_j)\left[(\bar{1}_{y_i} - \bar{\eta}_i)\cdot(\bar{1}_{y_j} - \bar{\eta}_j)\right] - \sum_i \bar{\eta}_i \cdot \bar{1}_{y_i}$$

$$\text{subject to}: \quad \forall i \; : \quad \bar{\eta}_i \geq 0 \quad \text{and} \quad \bar{\eta}_i \cdot \bar{1} = 1 \quad .$$

Решение задачи

- Произведем замену переменных $\bar{\tau}_i = \bar{1}_{y_i} - \bar{\eta}_i$
- Легко проверить, что $\mathcal{Q}(\tau)$ вогнут
- Следовательно есть максимум $\mathcal{Q}(\tau)$
- Тогда

$$\bar{M}_r = \beta^{-1}\left[\sum_i (\delta_{y_i,r} - \eta_{i,r})\bar{x}_i\right] \quad \Rightarrow \quad \bar{M}_r = \beta^{-1}\sum_i \tau_{i,r}\bar{x}_i$$

- Задача

$$\max_{\tau} \qquad \mathcal{Q}(\tau) = -\frac{1}{2}\sum_{i,j}(\bar{x}_i \cdot \bar{x}_j)(\bar{\tau}_i \cdot \bar{\tau}_j) + \beta\sum_i \bar{\tau}_i \cdot \bar{1}_{y_i}$$

$$\text{subject to :} \quad \forall i \quad \bar{\tau}_i \leq \bar{1}_{y_i} \quad \text{and} \quad \bar{\tau}_i \cdot \bar{1} = 0 \quad .$$

Финальный вид классификатора

$$H(\bar{x}) \quad = \quad \arg\max_{r=1}^{k} \{\bar{M}_r \cdot \bar{x}\} = \arg\max_{r=1}^{k} \left\{ \sum_i \tau_{i,r}(\bar{x}_i \cdot \bar{x}) \right\}$$

- Заметим, что двойственная функция и классификатор зависят только от $(\bar{x}_i \cdot \bar{x})$
- Следует можем перейти к ядровой функции $K(\cdot, \cdot)$
- Получим

$$H(\bar{x}) = \arg\max_{r=1}^{k} \left\{ \sum_i \tau_{i,r} K(\bar{x}, \bar{x}_i) \right\}$$

Двойственная задача

$$\max_{\tau} \qquad \mathcal{Q}(\tau) = -\frac{1}{2}\sum_{i,j} K\left(\bar{x}_i, \bar{x}_j\right)\ \left(\bar{\tau}_i \cdot \bar{\tau}_j\right) + \beta \sum_i \bar{\tau}_i \cdot \bar{1}_{y_i}$$

$$\text{subject to}: \quad \forall i \quad \bar{\tau}_i \leq \bar{1}_{y_i} \quad \text{and} \quad \bar{\tau}_i \cdot \bar{1} = 0 \quad ,$$

Нахождение решения двойственной задачи

- Задача может быть решена стандартными методами квадратичного программирования
- Поскольку используется m*k переменных, алгоритм переведет двойственную задачу в стандартный вид, где получим матрицу размером mk * mk
- Очевидно, что хранить матрицу такого размера невозможно для задач с большим количеством данных

Нахождение решения двойственной задачи

Решение авторов

- Основная идея разбить ограничения двойственной задачи на m непересекающихся множеств $\{\bar{\tau}_i | \bar{\tau}_i \leq \bar{1}_{y_i},\ \bar{\tau}_i \cdot \bar{1} = 0\}_{i=1}^{m}$
- Алгоритм работает пошагово, на каждой итерации выбирает паттерн p и увеличивает значение целевой функции, обновляя $\bar{\tau}_p$ с ограничениями $\bar{\tau}_p \leq \bar{1}_{y_p}$ и $\bar{\tau}_p \cdot \bar{1} = 0.$
- Поэтому на каждом шаге мы имеем k переменных и k + 1 ограничение

Нахождение решения двойственной задачи

Input $\{(\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m)\}$.
Initialize $\bar{\tau}_1 = \bar{0}, \ldots, \bar{\tau}_m = \bar{0}$.
Loop:

1. Choose an example p.
2. Calculate the constants for the reduced problem:
   - $A_p = K(\bar{x}_p, \bar{x}_p)$
   - $\bar{B}_p = \sum_{i \neq p} K(\bar{x}_i, \bar{x}_p)\bar{\tau}_i - \beta \bar{1}_{y_p}$
3. Set $\bar{\tau}_p$ to be the solution of the reduced problem :

$$\min_{\tau_p} \quad Q(\bar{\tau}_p) = \frac{1}{2} A_p (\bar{\tau}_p \cdot \bar{\tau}_p) + \bar{B}_p \cdot \bar{\tau}_p$$

$$\text{subject to :} \quad \bar{\tau}_p \leq \bar{1}_{y_p} \quad \text{and} \quad \bar{\tau}_p \cdot \bar{1} = 0$$

Output : $H(\bar{x}) = \arg\max_{r=1}^{k} \left\{ \sum_i \tau_{i,r} K(\bar{x}, \bar{x}_i) \right\}$.