

Машинное обучение

Лекция 5. Метод опорных векторов

Автор: Рустам Азимов

Санкт-Петербург, 2023г.

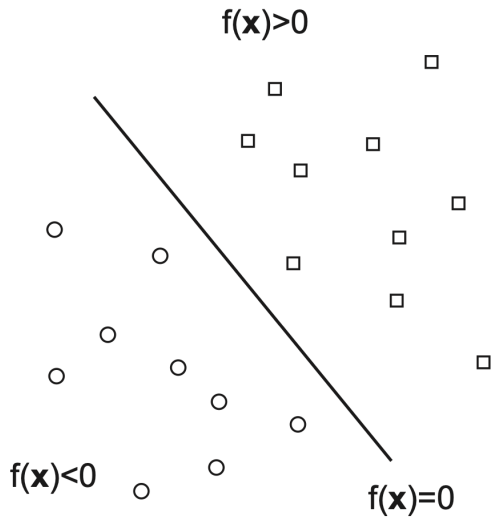
Задача бинарной классификации

- ▶ X — признаки, вещественные числа
- ▶ Целевой признак $Y \in \{-1, +1\}$
- ▶ Линейная модель для классификации:

$$f(x) = \text{sign}\left(\sum_{i=1}^m w_i x_i + b\right) = \text{sign}(\langle w, x \rangle + b)$$

- ▶ Разделяет пространство на две части гиперплоскостью
- ▶ Величина скалярного произведения описывает расстояние до гиперплоскости, а его знак — по какую сторону данный объект

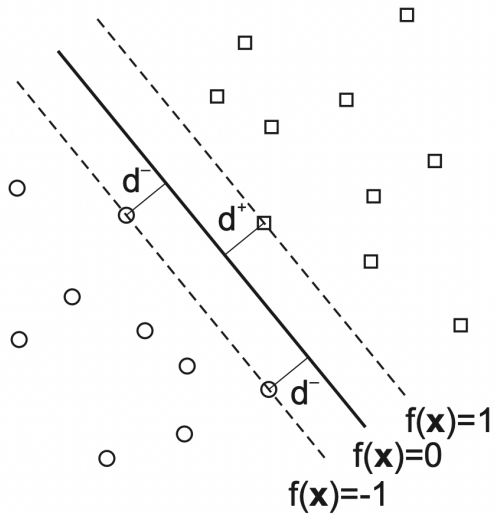
Разделяющая гиперплоскость



Разделимый случай

- ▶ Можно найти такие веса w , чтобы классификатор $f(x)$ не ошибался нигде на обучающей выборке
- ▶ В этом случае говорят, что выборка линейно разделима
- ▶ **Зазор (отступ, margin)** между классом и разделяющей гиперплоскостью — минимальное расстояние между гиперплоскостью и объектом класса
- ▶ Для наших двух классов обозначим зазор через d^+ и d^-

Зазоры



Оптимальная гиперплоскость

- ▶ В этом случае провести гиперплоскость, корректно разделяющую данные, можно разными способами
- ▶ Хотим хорошую обобщающую способность модели
- ▶ Тогда определим оптимальную гиперплоскость как гиперплоскость, максимизирующую зазор:

$$\min(d^+, d^-) \rightarrow \max_{w, b}$$

- ▶ Максимизация зазора между гиперплоскостью и данными позволяет надеяться на хорошую обобщающую способность в том случае, когда тестовая выборка является небольшой вариацией обучающей

- ▶ Очевидно, что при фиксированном направлении гиперплоскости (фиксированном векторе нормали w) оптимальная гиперплоскость проходит по середине между линиями уровня до ближайших объектов классов
- ▶ Таким образом, можно считать, что

$$d^+(w) = d^-(w)$$

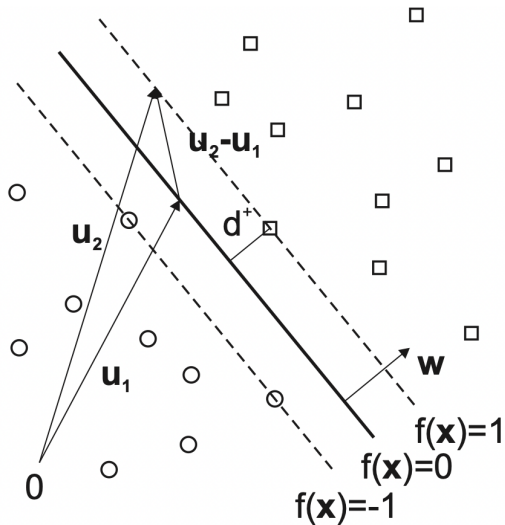
- ▶ Кроме того, гиперплоскость определена с точностью до масштаба шкалы измерения w и b
- ▶ Действительно, если умножить w и b на одно и тоже положительное число, то классификатор не изменится
- ▶ Зато перемещаются линии уровня $\langle w, x \rangle + b = a$
- ▶ Потребуем, чтобы линии уровня, проходящие через ближайшие объекты классов к гиперплоскости, определялись как $\langle w, x \rangle + b = 1$ и $\langle w, x \rangle + b = -1$

Величина зазоров

- ▶ Рассмотрим произвольный вектор u_1 , принадлежащий гиперплоскости, и произвольный вектор u_2 , принадлежащий линии уровня $\langle w, x \rangle + b = 1$
- ▶ Очевидно, что величина зазора d^+ равна длине проекции вектора $u_2 - u_1$ на вектор нормали w
- ▶ Таким образом, можно получить, что величина зазора

$$d^+ = d^- = \frac{1}{||w||}$$

Величина зазоров



Hard margin SVM

- ▶ Тогда запишем задачу максимизации зазора, которая и определяет **метод опорных векторов для линейно разделимой выборки (hard margin support vector machine)**

$$\left\{ \begin{array}{l} \frac{2}{||w||} \rightarrow \max_{w,b} \\ \langle w, x_i \rangle + b \geq 1, \quad \text{если } y_i = 1 \\ \langle w, x_i \rangle + b \leq -1, \quad \text{если } y_i = -1 \end{array} \right.$$

Hard margin SVM

- ▶ Тогда запишем задачу максимизации зазора, которая и определяет **метод опорных векторов для линейно разделимой выборки (hard margin support vector machine)**

$$\begin{cases} \frac{2}{\|w\|} \rightarrow \max_{w,b} \\ \langle w, x_i \rangle + b \geq 1, & \text{если } y_i = 1 \\ \langle w, x_i \rangle + b \leq -1, & \text{если } y_i = -1 \end{cases}$$

- ▶ Или эквивалентная система, но более удобная и с выпуклой функцией после добавления квадрата

$$\begin{cases} \frac{1}{2} \|w\|^2 \rightarrow \min_{w,b} \\ y_i \langle w, x_i \rangle + b \geq 1 \end{cases}$$

Неразделимый случай

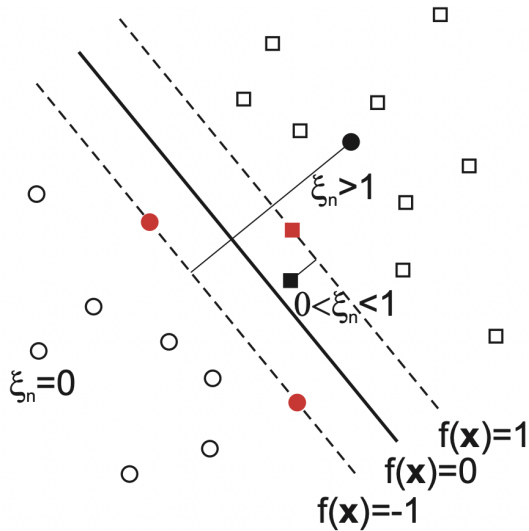
- ▶ Рассмотрим теперь случай произвольных данных, когда условие $y_i \langle w, x_i \rangle + b \geq 1$ не может быть выполнено для всех объектов обучающей выборки
- ▶ Тогда добавим в эти условия т.н. **ослабляющие коэффициенты** $\varepsilon_i \geq 0$:

$$y_i \langle w, x_i \rangle + b \geq 1 - \varepsilon_i$$

Ослабляющие коэффициенты

- ▶ $\varepsilon_i = 0$ — ошибки нет, объект x_i лежит за линиями уровня $|f(x)| = 1$
- ▶ $0 < \varepsilon_i \leq 1$ — ошибки нет, объект x_i лежит внутри корридора $0 \leq y_f(x) < 1$
- ▶ $\varepsilon_i > 1$ — ошибка есть, величина ошибки пропорциональна расстоянию от объекта x_i до гиперплоскости

Ослабляющие коэффициенты



- ▶ Добавим минимизацию ошибок в критерий оптимизации и получим критерий для метода опорных векторов в неразделимом случае (soft margin support vector machine)

$$\begin{cases} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \varepsilon_i \rightarrow \min_{w, b, \varepsilon_i} \\ y_i \langle w, x_i \rangle + b \geq 1 - \varepsilon_i \\ \varepsilon_i \geq 0 \end{cases}$$

- ▶ Здесь $C \geq 0$ — коэффициент регуляризации, определяющий компромисс между количеством ошибок и простотой модели (близость весов из w к нулю)
- ▶ C задается пользователем до начала обучения

Безусловная формулировка SVM

- ▶ Перейдем к безусловной оптимизационной задаче
- ▶ Перепишем условия

$$\begin{cases} \varepsilon_i \geq 1 - y_i(\langle w, x_i \rangle + b) \\ \varepsilon_i \geq 0 \end{cases}$$

- ▶ В форму, в которой ошибки как можно меньше

$$\varepsilon_i = \max(0, 1 - y_i(\langle w, x_i \rangle + b)) = l_{hinge}(y_i, f(x_i))$$

- ▶ Получим безусловную задачу для метода опорных векторов:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i(\langle w, x_i \rangle + b)) \rightarrow \min_{w, b, \varepsilon_i}$$

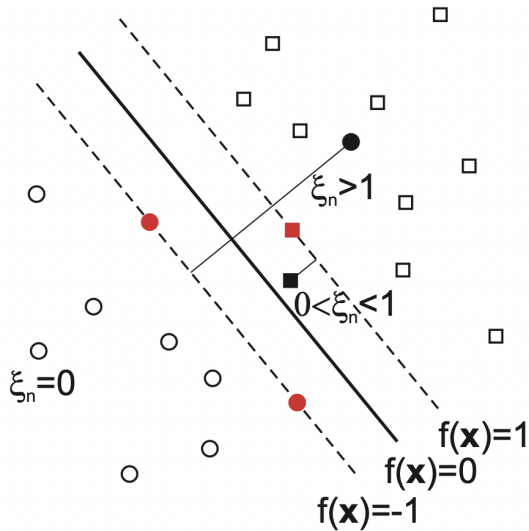
Двойственная задача для SVM

- ▶ Двойственная задача для SVM:

$$\begin{cases} -\frac{1}{2}\lambda^T \text{diag}(y)X^T X \text{diag}(y)\lambda + \lambda^T \vec{1} \rightarrow \max_{\lambda} \\ y^T \lambda = 0 \\ 0 \leq \lambda_i \leq C \end{cases}$$

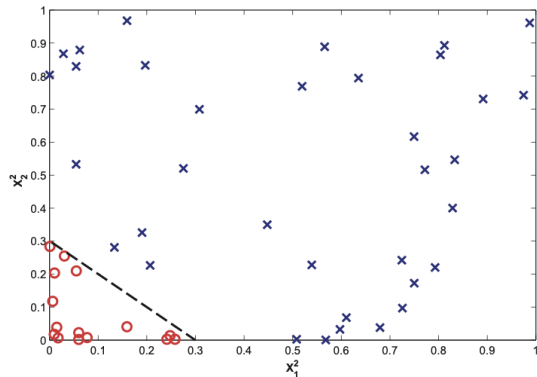
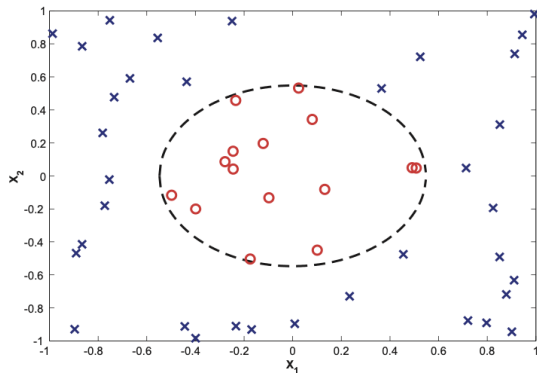
- ▶ Прямая задача оптимизации выпуклая, значит оптимальные значения функционалов прямой и двойственной задачи совпадают
- ▶ Более того, в данном случае решение прямой задачи (w^*, b^*) может быть выражено через решение двойственной задачи λ^*
- ▶ Получим классификатор $f(x) = (\sum_{i=1}^n \lambda_i^* y_i x_i^T)x + b^*$
- ▶ Где для "красных" объектов можно найти $b^* = y_i(1 - (w^*)^T x_i)$ (усреднив по этому правилу)

Опорные вектора



- ▶ На практике поверхность, разделяющая два класса, может быть существенно нелинейной
- ▶ Метод опорных векторов можно обобщить на случай построения нелинейных разделяющих поверхностей с помощью т.н. **ядрового перехода**

Ядровый переход к (x_1^2, x_2^2)



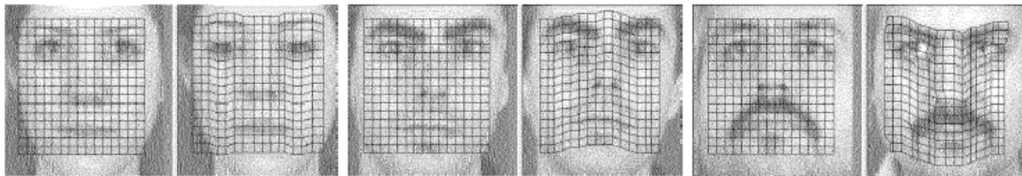
Ядровая функция

- ▶ Рассмотрим преобразование $\Phi : \mathbb{R} \rightarrow H$ из исходного пространства в новое
- ▶ Будем искать оптимальную разделяющую гиперплоскость в новом пространстве H с помощью метода опорных векторов
- ▶ Двойственная задача и решающая функция зависят только от скалярных произведений между объектами, поэтому достаточно знать лишь **ядровую функцию** для скалярных произведений:

$$\langle \Phi(x_1), \Phi(x_2) \rangle = K(x_1, x_2)$$

- ▶ Критерии для существования Φ :
 - ▶ Симметричность: $K(x_1, x_2) = K(x_2, x_1)$
 - ▶ Условие Мерсера
- ▶ Примеры ядровых функций:
 - ▶ Линейная: $K(x_1, x_2) = x_1^T x_2 + \theta, \theta \geq 0$
 - ▶ Степенная: $K(x_1, x_2) = (x_1^T x_2 + \theta)^d, \theta \geq 0, d \in \mathbb{N}$
 - ▶ Радиальная: $K(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}}, \sigma > 0$

- ▶ Использование ядерных функций также бывает оправдано в тех случаях, когда пространство объектов обладает сложной структурой (например, изображения)
- ▶ И задание скалярных произведений между парами объектов в таком пространстве оказывается легче, чем выбор пространства признаков
- ▶ Такой подход получил название беспризнакового распознавания образов



- ▶ http://www.machinelearning.ru/wiki/images/2/25/SMAIS11_SVM.pdf
- ▶ machinelearning.ru
- ▶ scikit-learn.org
- ▶ [kaggle](https://www.kaggle.com)