Обучение с учителем. Метод опорных векторов. Выбор модели с помощью кросс-валидации.

Лунев Иван, Высоков Максим, Петраков Михаил

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Кафедра статистического моделирования

Санкт-Петербург
2019г.

X — множество объектов, Y — множество ответов
$y : X \to Y$ — неизвестная зависимость (target function)

Дано: обучающая выборка — $(x_1, \dots, x_n) \subset X$, $y_i = y(x_i)$, $i = 1, \dots, n$ — известные ответы.
Найти: $a : X \to Y$ — функцию (decision function), приближающую y на всем множестве X.

Вероятностная постановка задачи: имеется неизвестное распределение на множестве $X \times Y$ с плотностью $p(x, y)$, из которого случайно выбираются $\mathbb{X}_n = (x_i, y_i)_{i=1}^n$ (независимые).

Задача классификации:

- $Y = \{-1, +1\}$ — классификация на 2 класса
- $Y = \{1, \dots, K\}$ — классификация на K классов

- Задача классификации с двумя классами, $Y = \{-1, +1\}$: по обучающей выборке $X^n = (x_i, y_i)_{i=1}^n$ построить алгоритм классификации $a(x, \mathrm{w}) = \operatorname{sign} f(x, \mathrm{w})$, где $f(x, \mathrm{w})$ — разделяющая (дискриминантная) функция, w — вектор параметров.

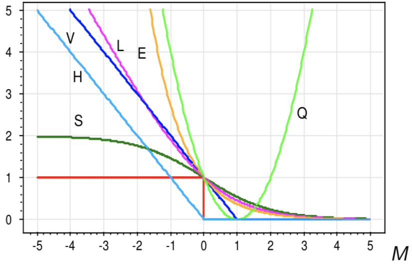
- $f(x, \mathrm{w}) = 0$ — разделяющая поверхность;
  $M_i(\mathrm{w}) = y_i f(x_i, \mathrm{w})$ — отступ (margin) объекта x_i;
  $M_i(\mathrm{w}) < 0 \Leftrightarrow$ алгоритм $a(x, \mathrm{w})$ ошибается на x_i.

- Минимизация эмпирического риска:

$$Q(\mathrm{w}) = \sum_{i=1}^{n} [M_i(\mathrm{w}) < 0] \leq \tilde{Q}(\mathrm{w}) = \sum_{i=1}^{n} \mathcal{L}(M_i(\mathrm{w})) \to \min_{\mathrm{w}};$$

функция потерь $\mathcal{L}(M_i(\mathrm{w}))$ невозрастающая, неотрицательная.

Часто используемые непрерывные функции потерь $\mathscr{L}(M)$:



$V(M) = (1 - M)_+$ — кусочно-линейная (SVM);
$H(M) = (-M)_+$ — кусочно-линейная (Hebb's rule);
$L(M) = \log_2(1 + e^{-M})$ — логарифмическая (LR);
$Q(M) = (1 - M)^2$ — квадратичная (FLD);
$S(M) = 2(1 + e^{M})^{-1}$ — сигмоидная (ANN);
$E(M) = e^{-M}$ — экспоненциальная (AdaBoost);
$[M < 0]$ — пороговая функция потерь.

Дано:

- Обучающая выборка $X^n = (x_i, y_i)_{i=1}^n$
- x_i — объекты, векторы из множества $X = \mathbb{R}^n$
- y_i — метки классов, элементы множества $Y = \{-1, 1\}$

Найти: Параметры $\mathrm{w} \in \mathbb{R}^p, \mathrm{w}_0 \in \mathbb{R}$ линейной модели классификации

$$a(x; \mathrm{w}, \mathrm{w}_0) = \mathrm{sign}(< x, \mathrm{w} > - \mathrm{w}_0).$$

Критерий — минимизация эмпирического риска:

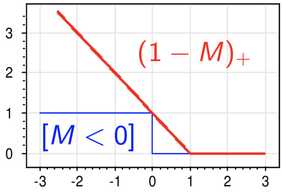
$$\sum_{i=1}^{n} [a(x_i; \mathrm{w}, \mathrm{w}_0) \neq y_i] = \sum_{i=1}^{n} [M_i(\mathrm{w}, \mathrm{w}_0) < 0] \to \min_{\mathrm{w}, \mathrm{w}_0},$$

где $M_i(\mathrm{w}, \mathrm{w}_0) = (< x, \mathrm{w} > - \mathrm{w}_0) y_i$ — отступ (margin) объекта x_i.

$$Q(\mathrm{w}, \mathrm{w}_0) = \sum_{i=1}^{n} [M_i(\mathrm{w}, \mathrm{w}_0) < 0] \leq \sum_{i=1}^{n} (1 - M_i(\mathrm{w}, \mathrm{w}_0))_+ + \frac{1}{2C} ||\,\mathrm{w}\,||^2 \to \min_{\mathrm{w}, \mathrm{w}_0}$$

Аппроксимация
штрафует объекты за приближение к границе классов, увеличивая зазор между классами

Регуляризация
штрафует неустойчивые решения в случае мультиколлинеарности



Линейный классификатор : $a(x, \mathrm{w}) = \mathrm{sign}(< x, \mathrm{w} > - \mathrm{w}_0)$
Пусть выборка $X^n = (x_i, y_i)_{i=1}^n$:

$$\exists\, \mathrm{w}, \mathrm{w}_0 : M_i(\mathrm{w}, \mathrm{w}_0) = (< x, \mathrm{w} > - \mathrm{w}_0) y_i > 0, i = 1 \dots n.$$

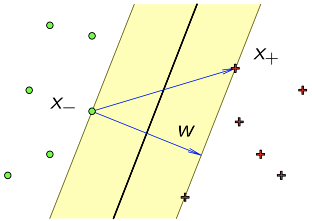
Нормировка: $\min\limits_{i=1,\dots,n} M_i(\mathrm{w}, \mathrm{w}_0) = 1.$

Разделяющая полоса (разделяющая гиперплоскость посередине):

$$x : -1 \leq < x, \mathrm{w} > - \mathrm{w}_0 \leq 1$$

$$\exists x_+ :< x_+, \mathrm{w} > - \mathrm{w}_0 = 1$$

$$\exists x_- :< x_-, \mathrm{w} > - \mathrm{w}_0 = -1$$



Ширина полосы:

$$\frac{< x_+ - x_-, \mathrm{w} >}{||\,\mathrm{w}\,||} = \frac{2}{||\,\mathrm{w}\,||} \to \max$$

Линейно разделимая выборка:

$$\begin{cases} \frac{1}{2}||\,\mathrm{w}\,||^2 \to \min\limits_{\mathrm{w},\mathrm{w}_0}; \\ M_i(\mathrm{w},\mathrm{w}_0) \geq 1, \quad i = 1,\dots,n \end{cases}$$

Переход к линейно неразделимой выборки:

$$\begin{cases} \frac{1}{2}||\,\mathrm{w}\,||^2 + C\sum_{i=1}^{n}\xi_i \to \min\limits_{\mathrm{w},\mathrm{w}_0,\xi}; \\ M_i(\mathrm{w},\mathrm{w}_0) \geq 1 - \xi_i, \qquad i = 1,\dots,n; \\ \xi_i \geq 0, \quad i = 1,\dots,n. \end{cases}$$

Эквивалентная задача безусловной минимизации:

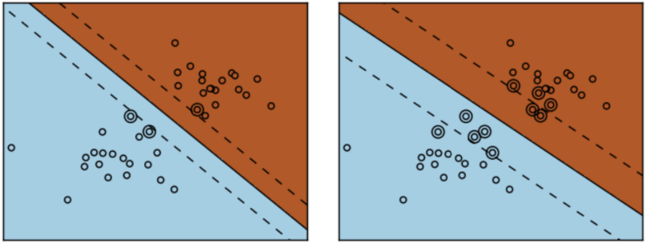
$$C\sum_{i=1}^{n}(1 - M_i(\mathrm{w},\mathrm{w}_0))_+ + \frac{1}{2}||\,\mathrm{w}\,||^2 \to \min_{\mathrm{w},\mathrm{w}_0,}.$$

SVM — аппроксимация и регуляризация эмпирического риска:

$$\sum_{i=1}^{n}(1 - M_i(\mathrm{w}, \mathrm{w}_0))_+ + \frac{1}{2C}||\,\mathrm{w}\,||^2 \to \min_{\mathrm{w},\mathrm{w}_0}.$$

большой C
слабая регуляризация

малый C
сильная регуляризация



Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_x \\ g_i(x) \leq 0, \quad i = 1, \ldots, m; \\ h_j(x) = 0, \quad i = 1, \ldots, k; \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \ldots, k$, λ_j, $j = 1, \ldots, k$:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, \quad \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{k} \lambda_j h_j(x); \\ g_j(x) \leq 0; h_j(x) = 0; \quad \text{(исходные ограничения)} \\ \mu_i \geq 0; \ \text{(двойственные ограничения)} \\ \mu_i g_j(x) = 0. \ \text{(условия дополняющей нежесткости)} \end{cases}$$

Функция Лагранжа:

$$\mathcal{L}(\mathrm{w}, \mathrm{w}_0, \xi; \lambda, \mu) = \frac{1}{2}||\,\mathrm{w}\,||^2 - \sum_{i=1}^{n} \lambda_i (M_i(\mathrm{w}, \mathrm{w}_0) - 1) - \sum_{i=1}^{n} \xi_i (\lambda_i + \mu_i - C),$$

λ_i — переменные, двойственные к ограничениям $M_i \geq 1 - \xi_i$;
μ_i — переменные, двойственные к ограничениям $\xi_i \geq 0$.

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial\,\mathrm{w}} = 0, \frac{\partial \mathcal{L}}{\partial\,\mathrm{w}_0} = 0, \frac{\partial \mathcal{L}}{\partial \xi} = 0; \\ \xi_i \geq 0, \quad \lambda_i \geq 0, \quad \mu_i \geq 0, \quad i = 1, \ldots, n; \\ \lambda_i = 0 \text{ либо } M_i(\mathrm{w}, \mathrm{w}_0) = 1 - \xi_i, \quad i = 1, \ldots, n; \\ \mu_i = 0 \text{ либо } \xi_i = 0, \quad i = 1, \ldots, n. \end{cases}$$

Функция Лагранжа:

$$\mathcal{L}(\mathrm{w}, \mathrm{w}_0, \xi; \lambda, \mu) = \frac{1}{2}||\,\mathrm{w}\,||^2 - \sum_{i=1}^{n} \lambda_i (M_i(\mathrm{w}, \mathrm{w}_0) - 1) - \sum_{i=1}^{n} \xi_i (\lambda_i + \mu_i - C),$$

Необходимые условия седловой точки Лагранжа:

$$\frac{\partial \mathcal{L}}{\partial\,\mathrm{w}} = \mathrm{w} - \sum_{i=1}^{n} \lambda_i y_i x_i = 0 \qquad \Longrightarrow \qquad w = \sum_{i=1}^{n} \lambda_i y_i x_i;$$

$$\frac{\partial \mathcal{L}}{\partial\,\mathrm{w}_0} = -\sum_{i=1}^{n} \lambda_i y_= 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{n} \lambda_i y_i = 0;$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = -\lambda_i - \mu_i + C = 0 \qquad \Longrightarrow \qquad \lambda_i + \mu_i = C, \quad i = 1, \ldots, n.$$

Типизация объектов:

- $\alpha_i = 0$; $\mu_i = C$; $\xi_i = 0$; $M_i \geq 1$ — периферийные (неинформативные) объекты;
- $0 < \alpha_i < C$; $0 < \mu_i < C$; $\xi_i = 0$; $M_i = 1$ — опорные граничные объекты;
- $\alpha_i = C$; $\mu_i = 0$; $\xi_i > 0$; $M_i < 1$ — опорные-нарушители.

Определение

Объект x_i называется опорным, если $\lambda_i \neq 0$.

Двойственная задача

$$
\begin{cases}
-\mathcal{L}(\lambda) = -\sum_{i=1}^{n} \lambda_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j < x_i, x_j > \to \min\limits_{\lambda}; \\
0 \leq \lambda_i \leq C \quad i = 1, \ldots, n; \\
\sum_{i=1}^{n} \lambda_i y_i = 0.
\end{cases}
$$

Решение прямой задачи выражается через решение двойственной:

$$
\begin{cases}
\mathsf{w} = \sum_{=1}^{n} \lambda_i y_i x_i; \\
\mathsf{w}_0 =< \mathsf{w}_0, x_i > -y_i, \text{ для любого i: } \lambda_i > 0, M_i = 1.
\end{cases}
$$

Линейный классификатор:

$$
a(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \lambda_i y_i < x_i, x > - \mathsf{w}_0 \right).
$$

Переход к спрямляющему пространству более высокой размерности: $\psi : X \to H$.

Определение

Функция $K : X \times X \to \mathbb{R}$ — ядро, если $K(x, x') =< \psi(x), \psi(x') >$ при некотором $\psi : X \to H$, где H — гильбертово пространство.

Теорема

Функция $K(x, x')$ является ядром тогда и только тогда, когда она симметрична: $K(x, x') = K(x', x)$; и неотрицательно определена:

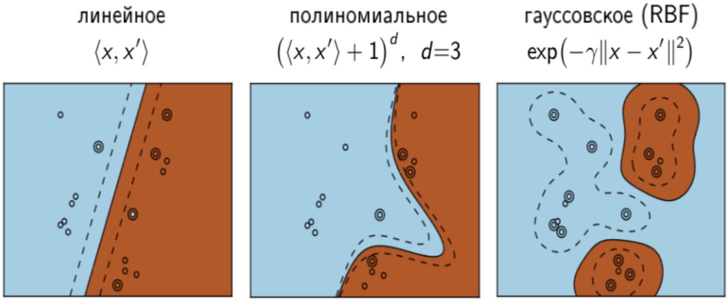
$$\int_X \int_X K(x, x')g(x)g(x')dxdx^{'} \geq 0 \text{ для людой } g : X \to \mathbb{R}.$$

1. $K(x,x') = < x, x' >$ — ядро;
2. константа $K(x,x') = 1$ — ядро;
3. произведение ядер $K(x,x') = K_1(x,x')K_2(x,x')$ — ядро;
4. $\forall \psi X \to \mathbb{R}$ произведение $K(x,x') = \psi(x), \psi(x')$ — ядро;
5. $K(x,x') = \alpha_1 K_1(x,x') + \alpha_2 K_2(x,x')$ при $\alpha_1, \alpha_2 > 0$ — ядро;
6. $\forall \phi : X \to X$ если K_0 ядро, то $K(x,x') = K_0(\phi(x), \phi(x'))$ — ядро;
7. если $s : X \times X \to \mathbb{R}$ — симметричная интегрируемая функция, то $K(x,x') = \int_X s(x,z)s(x',z)dz$ — ядро;
8. если K_0 — ядро и функция $f : \mathbb{R} \to \mathbb{R}$ представима в виде сходящегося степенного ряда с неотрицательными коэффициентами, то $K(x,x') = f(K_0(x,x'))$ — ядро.

1. $K(x,x') =< x,x' >^2$ — квадратичное ядро;
2. $K(x,x') =< x,x' >^d$ — полиномиальное ядро с мономами степени d;
3. $K(x,x') = (< x,x' > +1)^d$ — полиномиальное ядро с мономами степени $\leq d$;
4. $K(x,x') = \sigma(< x,x' >)$ — нейросеть с заданной функцией активации $\sigma(z)$ (не для всех σ является ядром);
5. $K(x,x') = th(k_1 < x,x' > -k_0), k_0, k_1 \geq 0$ — нейросеть с сигмоидными функциями активации;
6. $K(x,x') = exp(-\gamma||x - x'||^2)$ —сеть радиальных базисных функций (RBF ядро).

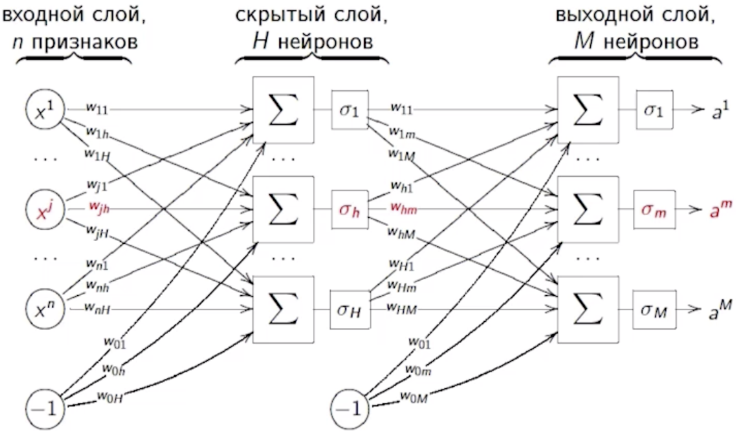
Гиперплоскость в спрямляющем пространстве соответствует нелинейной разделяющей поверхности в исходном.

Примеры с различными ядрами $K(x, x')$



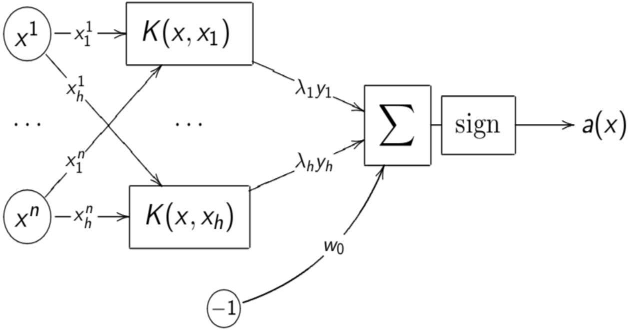
Двухслойная нейронная сеть

$$a^m(x) = \sigma_m\left(\sum_{h=0}^{H} w_{hm}\sigma_h\left(\sum_{j=0}^{J} w_{jh}f_j(x)\right)\right).$$



Перенумеруем объекты так, чтобы $x_1, \dots, x_h$ были опорными.

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{h} \lambda_i y_i K(x, x_i) - \mathrm{w}_0\right).$$



Первый слой вместо скалярных произведений вычисляет ядра.

Преимущества и недостатки SVM

Преимущества SVM:

- Задача выпуклого квадратичного программирования имеет единственное решение;
- Число нейронов скрытого слоя определяется автоматически — это число опорных векторов.

Недостатки SVM:

- Неустойчивость к шуму;
- Нет общий подходов к оптимизации $K(x, x')$ под задачу;
- Приходится подбирать константу C;
- Нет отбора признаков.

Дано:

- Имеется выборка (X_n, Y_n);
- Умеем строить модель, зависящую от параметра θ и минимизирующую ошибку $J(X_n, Y_n; \theta, \lambda)$, где λ — параметр регуляризации;

Хотим подобрать такой параметр θ_0, чтобы минимизировать ошибку $J(X_{new}, Y_{new}; \theta_0, 0)$ на новых индивидах.

Алгоритм:

- Делим выборку (X_n, Y_n) случайным образом на три набора: (X_{train}, Y_{train}), (X_{CV}, Y_{CV}) и (X_{test}, Y_{test}) ;
- Перебираем набор параметров $\lambda_1, \ldots, \lambda_m$;
- Для каждого параметра λ_i строим модель на (X$_{train}$, Y_{train}) (то есть находим оптимальное θ_{i0}) и считаем ошибку на $J(X_{CV}, Y_{CV}; \theta_{i0}, 0)$;
- Берем λ_0 с минимальной ошибкой (ему соответствует θ_0);
- Считаем ошибку модели $J(X_{test}, Y_{test}; \theta_0, 0)$.

K-fold Cross-validation

Условия такие же, как на предыдущем слайде.

Алгоритм:

- Делим выборку $(X_n,\, Y_n)$ случайным образом на K частей: $(X_1,\, Y_1)$,...,$(X_K,\, Y_K)$;
- Обозначим за $(X'_k,\, Y'_k)$ набор, содержащий всех индивидов, кроме $(X_k,\, Y_k)$;
- Перебираем набор параметров $\lambda_1, \ldots, \lambda_m$;
- Для каждого параметра λ_i считаем

$$CV_i = \sum_{j=1}^{K} \frac{n_j}{n} J(X_j, Y_j; \theta_j, 0),$$

  где θ_j минимизирует $J(X'_j, Y'_j; \theta, \lambda_i)$, n_j — число индивидов в $(X_j,\, Y_j)$;
- Берем λ_0 с минимальной ошибкой CV_i;
- Берем θ_0, которое минимизирует $J(X_n, Y_n; \theta, \lambda_0)$.