Обучение с учителем. Классификация. Дискриминантный анализ.

Е. Ларин, Ф. Ежов, И. Кононыхин

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Обучение с учителем

Выборка из генеральной случайной величины

- Для задачи регрессии: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$
- Для задачи классификации: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{A}^n$

Обучение с учителем: формальная постановка

- *Вход*: $\mathbf{X}$ — выборка $\boldsymbol{\xi}$, $\mathbf{y}$ — выборка η. Предполагаем, что существует неизвестное отображение $y^* : \boldsymbol{\xi} \to \eta$ (гипотеза непрерывности или компактности)
- *Задача*: По $\mathbf{X}$ и $\mathbf{y}$ найти такое отображение $\hat{y}^* : \boldsymbol{\xi} \to \eta$, которое приблизит отображение y^*.
- *Оценка*: Функция потерь $\mathfrak{L}(y^*(x), \hat{y}^*(x))$. Здесь x — реализация $\boldsymbol{\xi}$

Классификация

$$\mathbf{X} \in \mathbb{R}^{n \times p}, \quad \mathbf{y} \in \mathbb{A}^n \tag{1}$$

Гипотеза компактности

«Близкие» объекты, как правило, принадлежат одному классу

Понятие близости может быть формализовано, например, так:

$$\rho(\mathbf{x_1}, \mathbf{x_2}) = \left(\sum_{i=1}^{p} w_i |x_1^i - x_2^i|^k \right)^{\frac{1}{k}}$$

Классификация: генеральная постановка

Дано:

- $\boldsymbol{\xi} \in \mathbb{R}^p$ — вектор признаков
- $\eta \in \mathbb{A}$ — классовая принадлежность

Предположение об их зависимости можно записать в виде 2.

$$\eta = \Phi(\boldsymbol{\xi}) \tag{2}$$

Задача: найти Φ

Классификация: выборочная постановка

Дано:

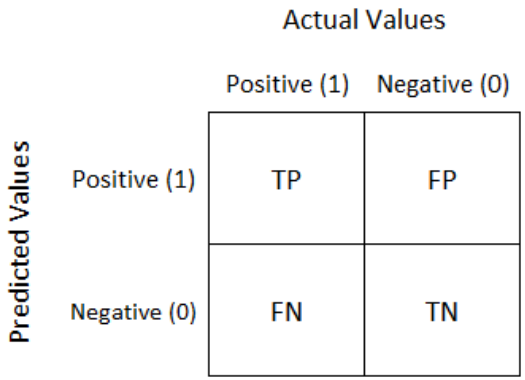
- $\mathbf{X} \in \mathbb{R}^{n \times p}$ — матрица признаков
- $\mathbf{y} \in \mathbb{A}^n$ — вектор классовой принадлежности

Предположение имеет вид 3.

$$y_i = \Phi(\mathbf{x}_i), \quad i = 1, \ldots, n \tag{3}$$

Задача: найти Φ

Классификация: оценка качества



На основе этой матрицы есть большое количество разных метрик: *accuracy*, *recall*, *precision*, F_β, *ROC–AUC*

Классификация: типы классов

- По количеству классов:
  - бинарная классификация
  - многоклассовая классификация
- По пересечению классов
  - пересекающиеся
  - непересекающаяся
  - нечёткие

Классификация: этапы обучения модели

- Выбор модели (класс рассматриваемых Φ из 3)
- Выбор метрики
- Выбор метода обучения (способ подбора параметров для минимизации метрики на обучающем множестве)
- Выбор метода проверки (способ оценки качества модели)

Классификация: задача оптимизации

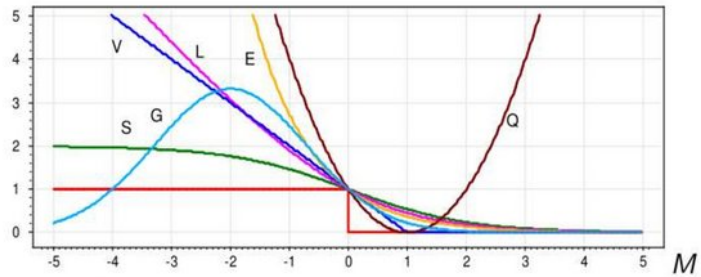
- $\hat{\beta}$ — параметры модели
- $\mathbf{\Phi}(\mathbf{x}, \beta)$ — функционал классификации
- $\mathfrak{L}(\mathbf{\Phi}(\mathbf{x}, \beta), \mathbf{y})$ — функция потерь (метрика)

$$\hat{\beta} = \arg\min_{\beta} \mathfrak{L}(\mathbf{\Phi}(\mathbf{x}, \beta), \mathbf{y})$$

Классификация: отступы

Можно ввести ту же задачу через оптимизацию функции от отступов.

Функции потерь $\mathscr{L}(M)$ в задачах классификации на два класса



$E(M) = e^{-M}$ — экспоненциальная (AdaBoost);
$L(M) = \log_2(1 + e^{-M})$ — логарифмическая (LogitBoost);
$G(M) = \exp(-cM(M + s))$ — гауссовская (BrownBoost);
$Q(M) = (1 - M)^2$ — квадратичная;
$S(M) = 2(1 + e^{M})^{-1}$ — сигмоидная;
$V(M) = (1 - M)_+$ — кусочно-линейная (SVM);

Классификация: общий подход к решению

Как построить функционал Ф?
Общий подход — построить набор f_i, $i = 1, \ldots, K$. Каждая функция $f_i(\mathbf{x})$ показывает меру принадлежности $\mathbf{x}$ классу i.
Таким образом,

$$\Phi(\mathbf{x}) = \arg\max_i(f_i(\mathbf{x})). \tag{4}$$

Дискриминантный анализ

Примем за функции f_i из 4 оценку вероятности принадлежности к i-му классу.

$$\Phi(\mathbf{x}) = \arg\max_i(P(C_i|\mathbf{x})).$$

C_i — класс, состоящий из одного события: $\mathbf{x}$ принадлежит i-му классу.

Дискриминантный анализ

Если известны априорные вероятности получения i-го класса (π_i), применим формулу Байеса

$$P(C_i|\mathbf{x}) = \frac{\pi_i P(\mathbf{x}|C_i)}{\sum_{j=1}^{K} \pi_j P(\mathbf{x}|C_j)}.$$

Отбросим знаменатель

$$f_i = P(C_i|\mathbf{x}) = \pi_i P(\mathbf{x}|C_i).$$

LDA

Предположение:

$$\mathrm{P}(\boldsymbol{\xi}|\eta = A_i) = \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$$

Классифицирующая функция:

$$f_i(\mathbf{x}) = \frac{\pi_i}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\right)$$

После упрощения:

$$h_i(\mathbf{x}) = -0.5\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i^{\mathrm{T}} + \boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\mathbf{x} + \log\pi_i$$

QDA

Предположение:

$$\mathrm{P}(\boldsymbol{\xi}|\eta = A_i) = \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Классифицирующая функция:

$$f_i(\mathbf{x}) = \frac{\pi_i}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\right)$$

После упрощения:

$$g_i(\mathbf{x}) = -0.5(\mathbf{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} - 0.5\log|\boldsymbol{\Sigma}_i| + \log\pi_i$$

Немного про регрессию

Обучающая выборка: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$.

1. Модель регрессии:

$$\hat{\mathbf{y}} = \mathbf{\Phi}(\mathbf{x}, \beta) = \langle \mathbf{x}, \beta \rangle = \sum_{j=1}^{p} \beta_j x_j, \ \beta \in \mathbb{R}^p$$

2. Функция потерь:

$$\mathfrak{L}(\hat{\mathbf{y}}, \mathbf{y}) = (\hat{\mathbf{y}} - \mathbf{y})^2$$

3. Метод обучения — метод наименьших квадратов:

$$Q(\beta) = \sum_{i=1}^{n} (\mathbf{\Phi}(\mathbf{x}_i, \beta) - \mathbf{y}_i)^2 \to \min_{\beta}$$

Со стороны классификации

Обучающая выборка: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $y \in \{-1, 1\}$.

1. Модель классификации:

$$\hat{y} = \mathbf{\Phi}(\mathbf{x}, \beta) = sign\langle \mathbf{x}, \beta \rangle = sign \sum_{j=1}^{p} \beta_j x_j, \ \beta \in \mathbb{R}^p$$

2. Функция потерь:

$$\mathfrak{L}(\hat{y}, y) = [\hat{y}y < 0] = [\langle \mathbf{x}, \beta \rangle y < 0] \leqslant \hat{\mathfrak{L}}(\langle \mathbf{x}, \beta \rangle y)$$

3. Метод обучения — минимизация эмпирического риска:

$$Q(\beta) = \sum_{i=1}^{n} [\langle \mathbf{x}_i, \beta \rangle y_i < 0] \leqslant \sum_{i=1}^{n} \hat{\mathfrak{L}}(\langle \mathbf{x}_i, \beta \rangle y_i) \to \min_{\beta}$$

Отступы

$\mathbf{\Phi}(\mathbf{x}, \beta) = sign(g(\mathbf{x}, \beta))$ — разделяющий классификатор,
$g(\mathbf{x}, \beta)$ — разделяющая функция,
$g(\mathbf{x}, \beta) = 0$ — уравнение разделяющей поверхности.

$M_i(\beta) = g(\mathbf{x}_i, \beta) y_i$ — отступ объекта $\mathbf{x}_i$.
Если $M_i(\beta) < 0$, тогда алгоритм ошибается на $\mathbf{x}_i$.

Отступы

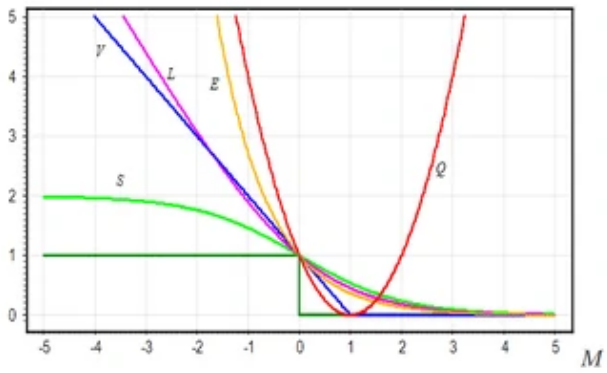


Рис. 7. Непрерывные аппроксимации пороговой функции потерь $[M < 0]$.

$Q(M) = (1 - M)^2$ — квадратичная;
$V(M) = (1 - M)_+$ — кусочно-линейная;
$S(M) = 2(1 + e^M)^{-1}$ — сигмоидная;
$L(M) = \log_2(1 + e^{-M})$ — логистическая;
$E(M) = e^{-M}$ — экспоненциальная.

Логистическая регрессия

Линейная модель классификации для двух классов $y \in \{-1, 1\}$:

$$\mathbf{\Phi}(\mathbf{x}, \beta) = sign(\langle \mathbf{x}, \beta \rangle)$$

Отступ $M_i = \langle \mathbf{x}_i, \beta \rangle y_i$.
Логистическая функция потерь: $\mathfrak{L}(M) = \log(1 + e^{-M})$
Модель условной вероятности: $P(y|\mathbf{x}, \beta) = \sigma(M) = \dfrac{1}{1 + e^{-M}}$

Логистическая регрессия: SoftMax

Линейный классификатор при произвольном числе классов $|Y|$:

$$\mathbf{\Phi}(\mathbf{x}, \beta) = \arg\max_{y \in Y} \langle \mathbf{x}, \beta_y \rangle$$

Вероятность того, что объект $\mathbf{x}$ относится к классу y:

$$P(y|\mathbf{x}, \beta) = \frac{e^{\langle \mathbf{x}, \beta_y \rangle}}{\sum\limits_{z \in Y} e^{\langle \mathbf{x}, \beta_z \rangle}} = SoftMax_{y \in Y} \langle \mathbf{x}, \beta_y \rangle$$

Функция $SoftMax : \mathbb{R}^Y \to \mathbb{R}^Y$ переводит произвольный вектор в нормированный вектор дискретного распределения.

Логистическая регрессия: Вероятностная постановка задачи

Зададим модель логистической регрессии следующим образом:

$$\log \frac{P(\eta = G_i|\boldsymbol{\xi} = \mathbf{x})}{P(\eta = G_K|\boldsymbol{\xi} = \mathbf{x})} = \beta_{i0} + \beta_i^T \mathbf{x}, i = 1, \cdots, K-1.$$

Перейдем от логитов к вероятностям:

$$P(\eta = G_i|\boldsymbol{\xi} = \mathbf{x}) = \frac{e^{\beta_{i0}+\beta_i^T \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0}+\beta_k^T \mathbf{x}}}, i = 1, \cdots, K-1,$$

$$P(\eta = G_K|\boldsymbol{\xi} = \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0}+\beta_k^T \mathbf{x}}}.$$

Логистическая регрессия: Вероятностная постановка задачи

Для оценки параметров воспользуемся методом максимального правдоподобия:

$$l(\theta) = \sum_{i=1}^{N} \log P(\eta = G_k | \boldsymbol{\xi} = \mathbf{x}; \theta),$$

$$\theta = (\beta_{10}, \boldsymbol{\beta}_1^T, \ldots, \beta_{(K-1)0}, \boldsymbol{\beta}_{K-1}^T).$$

Iteratively reweighted least squares (IRLS).

Функция потерь: $\mathfrak{L}(M_i(\boldsymbol{\beta})) = \log(1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i})$

SVM: Постановка задачи

Выборка: $\{x_i, y_i\}_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$.
Задача построить классифицирующие правило.

Предположим, что данные - разделимы гиперплоскостью,

$$\mathbf{x}^T\beta - \beta_0 = 0; \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R},$$

$$g(x) = \mathbf{x}^T\beta - \beta_0,$$

$$\mathbf{\Phi}(x) = sign(g(x)).$$

SVM: Постановка задачи

Критерий оптимальности: максимальное расстояние между двумя гиперплоскостями, параллельных данной и симметрично расположенных относительно нее.

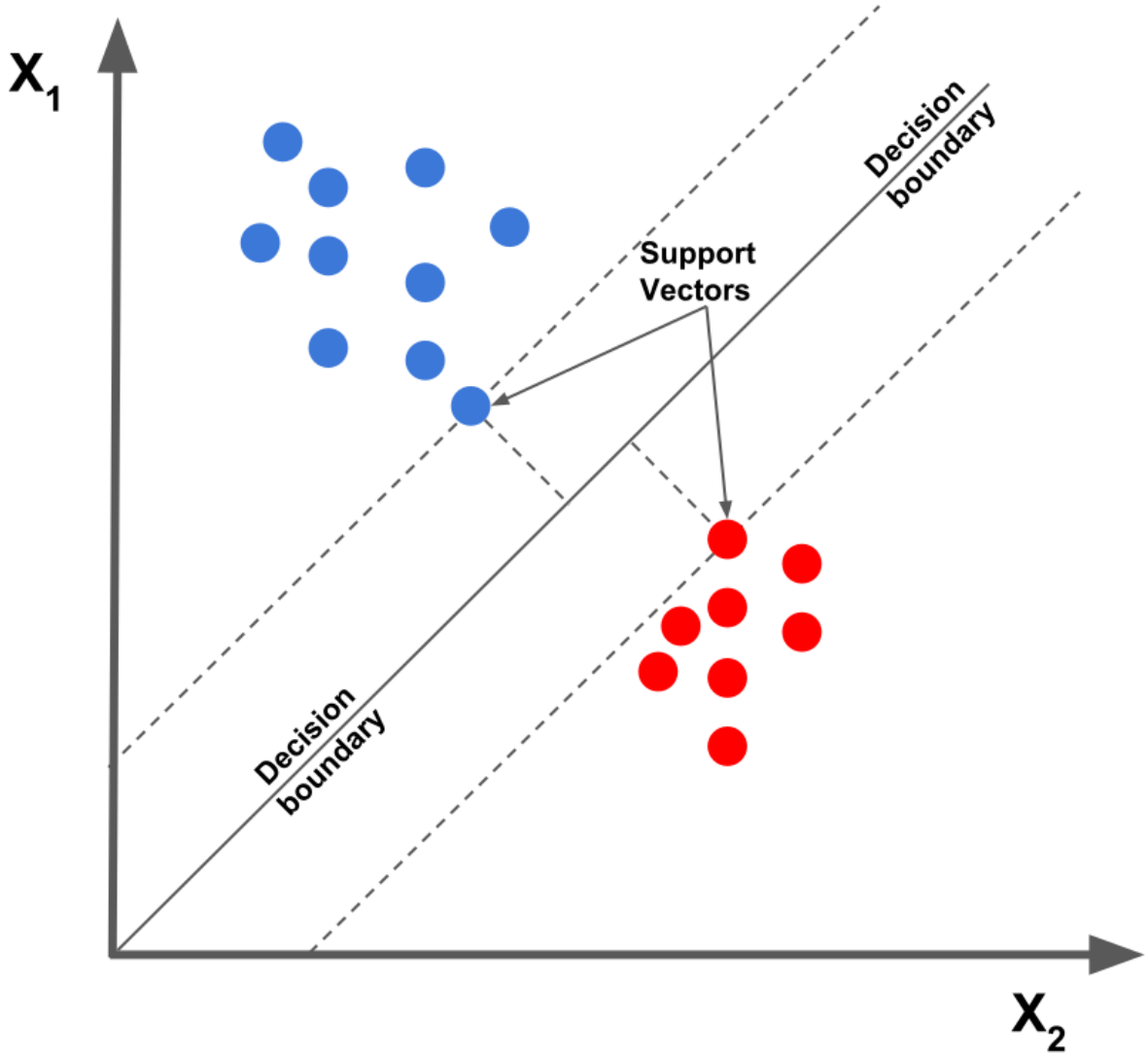
Эта пара гиперплоскостей может быть описана парой уравнений:

$$\mathbf{x}^T\beta - \beta_0 = -1,$$

$$\mathbf{x}^T\beta - \beta_0 = 1.$$

Растояние между ними: $\frac{2}{||\beta||}$.

SVM: Постановка задачи



SVM: Задача квадратичного программирования

Принадлежность точек обучающей выборки полупространства описывается

$$M_i = (\mathbf{x}^T\beta - \beta_0)y_i \geqslant 1$$

Задача сводится к задаче квадратичного программирования с линейными ограничениями:

$$\begin{cases} \frac{1}{2}||\beta||_2^2 \to \min\limits_{\beta,\beta_0} \\ M_i \geqslant 1 \end{cases}$$

Случай когда нету линейной разделимости:

$$\begin{cases} \frac{1}{2}||\beta||_2^2 + C\sum\xi_i \to \min\limits_{\beta,\beta_0,\xi} \\ M_i \geqslant 1 - \xi_i \\ \xi_i \geqslant 0 \end{cases}$$

SVM: Условия Каруша-Куна-Таккера

Применение условий Каруша-Куна-Таккера к задаче SVM

Функция Лагранжа:

$$L(\beta, \beta_0, \xi; \lambda, \eta) = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n}(M_i - 1) - \sum_{i=1}^{n}\xi_i(\lambda_i + \eta_i - C),$$

λ_i — переменные, двойственные к ограничениям $M_i \geqslant 1 - \xi_i$,

η_i — переменные, двойственные к ограничениям $\xi_i \geqslant 0$.

$$\begin{cases} \dfrac{\partial L}{\partial \beta} = 0, \dfrac{\partial L}{\partial \beta_0} = 0, \dfrac{\partial L}{\partial \xi} = 0; \\ \xi_i \geqslant 0, \lambda_i \geqslant 0, \eta_i \geqslant 0, i = 1, \ldots, n; \\ \lambda_i = 0 \text{ либо } M_i = 1 - \xi_i, i = 1, \ldots, n; \\ \eta_i = 0 \text{ либо } \xi_i = 0, i = 1, \ldots, n; \end{cases}$$

SVM

Необходимые условия седловой точки функции Лагранжа:

$$\frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i = 0 \Longrightarrow \beta = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial \beta_0} = -\sum_{i=1}^{n} \lambda_i y_i = 0 \Longrightarrow \sum_{i=1}^{n} \lambda_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = \lambda_i - \eta_i - C = 0 \Longrightarrow \lambda_i + \eta_i = C, i = 1, \ldots, n;$$

SVM: Опорные объекты

Типизация объектов:

1. $\lambda_i = 0; \eta_i = C; \xi = 0; M_i \geqslant 1$ - неинформативные объекты.
2. $0 < \lambda_i < C; 0 < \eta_i < C; \xi = 0; M_i = 1$ - опорные граничные объекты.
3. $\lambda_i = C; \eta_i = 0; \xi > 0; M_i < 1$ - опорные-нарушители.

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

SVM

Решаем:

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

Подставим полученные λ:

$$\begin{cases} \beta = \sum\limits_{i=1}^{n} \lambda_i y_i x_i; \\ \beta_0 = \langle \beta, \mathbf{x}_i \rangle - y_i, \text{ для любого } i : \lambda_i > 0, M_i = 1. \end{cases}$$

Линейный классификатор: $\mathbf{\Phi}(x) = sign(\sum\limits_{i=1}^{n} \lambda_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle - \beta_0)$

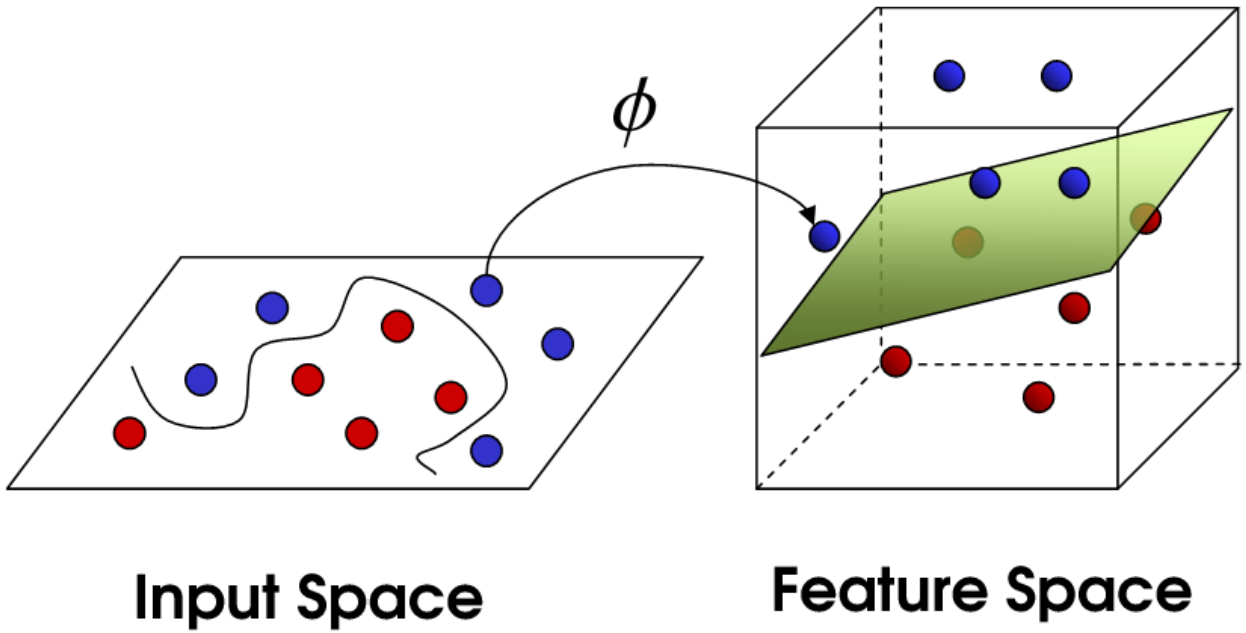
SVM: Нелинейное обобщение

Заменим $\langle \mathbf{x}, \mathbf{x}' \rangle$ нелинейной функцией $K(\mathbf{x}, \mathbf{x}')$.

Определение

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

SVM: Нелинейное обобщение



SVM: Примеры ядер

1. $K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^2$ — квадратичное ядро.
2. $K(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle^d$ — полиномиальное ядро степени d.
3. $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma ||\mathbf{x} - \mathbf{x}'||^2)$ — сеть радиальных базисных функций (RBF ядро).

Кросс-валидация

Кросс-валидация (aka перекрестная проверка, скользящий контроль) — процедура эмпирического оценивания обобщающей способности алгоритмов.
С помощью кросс-валидации "эмулируется"наличие тестовой выборки, которая не участвует в обучении модели, но для которой известны правильные ответы.

Кросс-валидация: виды

- Валидация на отложенных данных (Hold-Out Validation);
- Полная кросс-валидация (Complete Cross-Validation);
- k-fold Cross-Validation;
- t×k-fold Cross Validation;
- Кросс-валидация по отдельным объектам (Leave-One-Out);
- Случайные разбиения (Random Subsampling).

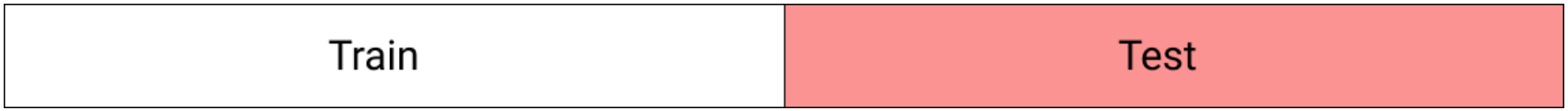
Кросс-валидация: Обозначения

Введем обозначения:

- $\mathbf{X}$ — матрица признаков, описывающие таргеты $\mathbf{y}$;
- $\mathbf{T}^N = (x_i, y_i)_{i=1}^N, x_i \in \mathbf{X}, y_i \in \mathbf{y}$ — обучающая выборка;
- $\mathfrak{L}$ — функция потерь (мера качества);
- A — исследуемая модель;
- $\mu : (\mathbf{X} \times \mathbf{y}) \to A$ — алгоритм обучения.

Кросс-валидация: Hold-Out Validation

Обучающая выборка один раз случайным образом разбивается на две части $\mathbf{T} = \mathbf{T}^t \cup \mathbf{T}^{N-t}$.



После чего решается задача оптимизации:
$HO(\mu, \mathbf{T}^t, \mathbf{T}^{N-t}) = \mathfrak{L}(\mu(\mathbf{T}^t), \mathbf{T}^{N-t}) \to min.$
Метод Hold-out применяется в случаях больших датасетов, т.к. требует меньше вычислительных мощностей по сравнению с другими методами кросс-валидации.
Недостатком метода является то, что оценка существенно зависит от разбиения, тогда как желательно, чтобы она характеризовала только алгоритм обучения.

Кросс-валидация: Complite cross-validation

- Выбирается значение t;
- Выборка разбивается всевозможными способами на две части $\mathbf{T} = \mathbf{T}^t \cup \mathbf{T}^{N-t}$.

После чего решается задача оптимизации:

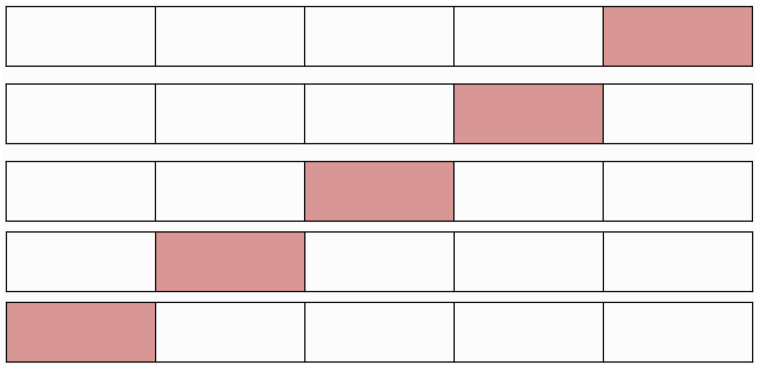
$$CVV_t = \frac{1}{C_N^{N-t}} \sum_{\mathbf{T}^N = \mathbf{T}^t \cup \mathbf{T}^{N-t}} \mathfrak{L}(\mu(\mathbf{T}^t), \mathbf{T}^{N-t}) \to min.$$

Здесь число разбиений C_N^{N-t} становится слишком большим даже при сравнительно малых значениях t, что затрудняет практическое применение данного метода.

Кросс-валидация: k-fold Cross-Validation

- $\mathbf{T}^l$ разбивается на $\mathbf{T_1} \cup \cdots \cup \mathbf{T_k}$, $|\mathbf{T_i}| \approx \frac{1}{k}$ частей;
- Производится k итераций:
  - Модель обучается на $k-1$ части обучающей выборки;
  - Модель тестируется на части обучающей выборки, которая не участвовала в обучении.

Каждая из k частей единожды используется для тестирования.



После чего решается задача оптимизации:

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} \mathfrak{L}(\mu(\mathbf{T}^N \setminus \mathbf{T}_i), \mathbf{T}_i) \to min.$$

Кросс-валидация: t×k-fold Cross Validation

Как k-fold Cross-Validation, только t раз.

Разбиение: $\mathbf{T}^N = \mathbf{T}_{(1,1)} \cup \cdots \cup \mathbf{T}_{(k,1)} = \mathbf{T}_{(1,t)} \cup \cdots \cup \mathbf{T}_{(k,t)}, |\mathbf{T}_{(i,j)}| \approx \frac{N}{k},$

Задача оптимизации: $CV_{t\times k} = \frac{1}{tk} \sum_{j=1}^{t} \sum_{i=1}^{k} \mathfrak{L}(\mu(\mathbf{T}^N \setminus \mathbf{T}_{(i,j)}), \mathbf{T}_{(i,j)}) \to min$

Кросс-валидация: Leave-One-Out

Выборка разбивается на $N-1$ и 1 объект N раз.

$LOO = \frac{1}{N}\sum_{i=1}^{N} \mathfrak{L}(\mu(\mathbf{T}^N \setminus p_i), p_i) \to min$, где $p_i = (x_i, y_i)$.

Преимущества LOO в том, что каждый объект ровно один раз участвует в контроле, а длина обучающих подвыборок лишь на единицу меньше длины полной выборки.

Недостатком LOO является большая ресурсоёмкость, так как обучаться приходится N раз.

Кросс-валидация: Random subsampling (Monte-Carlo cross-validation)

Выборка разбивается в случайной пропорции. Процедура повторяется несколько раз.

Кросс-валидация: Выбор лучшей модели

Не переобученый алгоритм должен показывать одинаковую эффективность на каждой части.

SGD

Стохастический градиентный спуск - оптимизационный алгоритм, при котором градиент оптимизируемой функции считается на каждой итерации от случайно выбранного элемента выборки.

Вспомним про GD: обозначения

- $\mathbf{X} \in \mathrm{R}^{n \times p}$ — матрица признаков;
- $y^* : \mathbf{X} \to \mathbf{y}$ - целевая зависимость, известная на $\mathbf{X}^l$;
- $\mathbf{y} = y^*(\mathbf{X}) \in \mathrm{A}^n$ — вектор классовой принадлежности;
- $\mathbf{X}^l$ — обучающая выборка, состоящая из пар $(x_i, y_i)_{i=1}^l$;
- $a(x, \omega)$ — семейство алгоритмов с параметром вектора весов ω;
- $\mathfrak{L}(a, y)$ — функция потерь, выбранная из класса C_1.

Вспомним про GD: постановка задачи

Найдем алгоритм $a(x, \omega)$, аппроксимирующий зависимость y^*.
Например, в случае линейного классификатора искомый алгоритм имеет вид:

$$a(x, \omega) = \phi(\sum_{j=1}^{N} \omega_j x^j - \omega_0),$$

где $\phi(z)$ - функция активации.
Решаемая задача оптимизации:

$$Q(\omega) = \frac{1}{N} \sum_{i=1}^{N} \mathfrak{L}(a(x_i, \omega), y_i) \to \min_{\omega}$$

Обновление весов:

$$\omega^{(t+1)} = \omega^{(t)} - h \nabla Q(\omega^{(t)})$$

Вспомним про GD: Проблемы

Проблема GD — чтобы определить новое приближение вектора весов необходимо вычислить градиент от каждого элемента выборки, что может сильно замедлять алгоритм. Идея ускорения заключается в использовании либо одного элемента (SGD), либо некоторой подвыборки (Batch GD) для получения нового приближения весов.

Ускоряем GD - SGD

Веса, при подходе SGD, обновляются следующим образом:

$$\omega^{(t+1)} = \omega^{(t)} - h\nabla\mathfrak{L}(a(x_i, \omega^{(t)}), y_i),$$

где i — случайно выбранный индекс.
Т.к. направление изменения ω будет определяться за $O(1)$, подсчет Q на каждом шаге будет слишком дорогостоящим. Для ускорения оценки, будем использовать одну из рекуррентных формул:

- Среднее арифметическое:
  $\overline{Q}_m = \frac{1}{m}\mathfrak{L}(a(x_{i_m}, \omega^{(m)}), y_{i_m}) + (1 - \frac{1}{m})\overline{Q}_{m-1}$;
- Экспоненциальное скользящее среднее:
  $\overline{Q}_m = \lambda\mathfrak{L}(a(x_{i_m}, \omega^{(m)}), y_{i_m}) + (1 - \lambda)\overline{Q}_{m-1}$, где λ — темп забывания "предыстории"ряда.

SGD: А что с инициализацией весов?

Существует несколько способов:

- $\omega = 0$;
- $\omega_j = random(-\frac{1}{2n}, \frac{1}{2n})$;
- $\omega_j = \frac{\langle y, x_j \rangle}{\langle x_j, x_j \rangle}$, если признаки независимы, функция активации линейна, а функция потерь квадратична.

SGD: А что со сходимостью?

Установлено, что если скорость обучения убывает при увеличении числа итераций SGD сходится почти наверняка к глобальному минимуму в случае выпуклой или псевдовыпуклой функции $Q(\omega)$, в противном случае метод сходится почти наверняка к локальному минимуму.

SGD: Достоинства и недостатки

Достоинства:

- Легко реализуется;
- Функция потерь и семейство алгоритмов могут быть любыми (если функция потерь не дифференцируема, ее можно аппроксимировать дифференцируемой);
- Легко добавить регуляризацию;
- Возможно потоковое обучение;
- Подходит для задач с большими данными, иногда можно получить решение даже не обработав всю выборку.

Недостатки:

- Нет универсального набора эвристик, их нужно выбирать для конкретной задачи отдельно;
- На практике остаются проблемы с локальными экстремумами.

Модификации (или о чем можно почитать на досуге)

- Не выраженные явно изменения (ISGD) — градиент пересчитывается на следующей итерации;
- Метод импульса — запоминает Δw на каждой итерации и определяет следующее изменение в виде линейной комбинации градиента и предыдущего изменения;
- Усреднённый стохастический градиентный спуск;
- AdaGrad — модификация SGD с отдельной для каждого параметра скоростью обучения;
- RMSProp — это метод, в котором скорость обучения настраивается для каждого параметра. Идея заключается в делении скорости обучения для весов на скользящие средние значения недавних градиентов для этого веса;
- Adam — это обновление оптимизатора RMSProp. Здесь используются скользящие средние как градиентов, так и вторых моментов градиентов.
- Kalman-based SGD.