# SVM

Д. Корчемкин, В. Агеев
622 группа

26 ноября 2017 г.

# SVM

С помощью SVM будет решаться задача классификации
Входные данные: выборка $\{x_i, y_i\}_{i=1}^n$ $x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$.
Задача: построить классифицирующее правило

$$h : \mathbb{R}^p \to \{-1, 1\}$$

такое, что $y_i \sim h(x_i)$ в некотором смысле.

# Hard-margin SVM

Предположим, что данные – разделимы гиперплоскостью

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

Определив величину, пропорциональную расстоянию до гиперплоскости (с знаком)

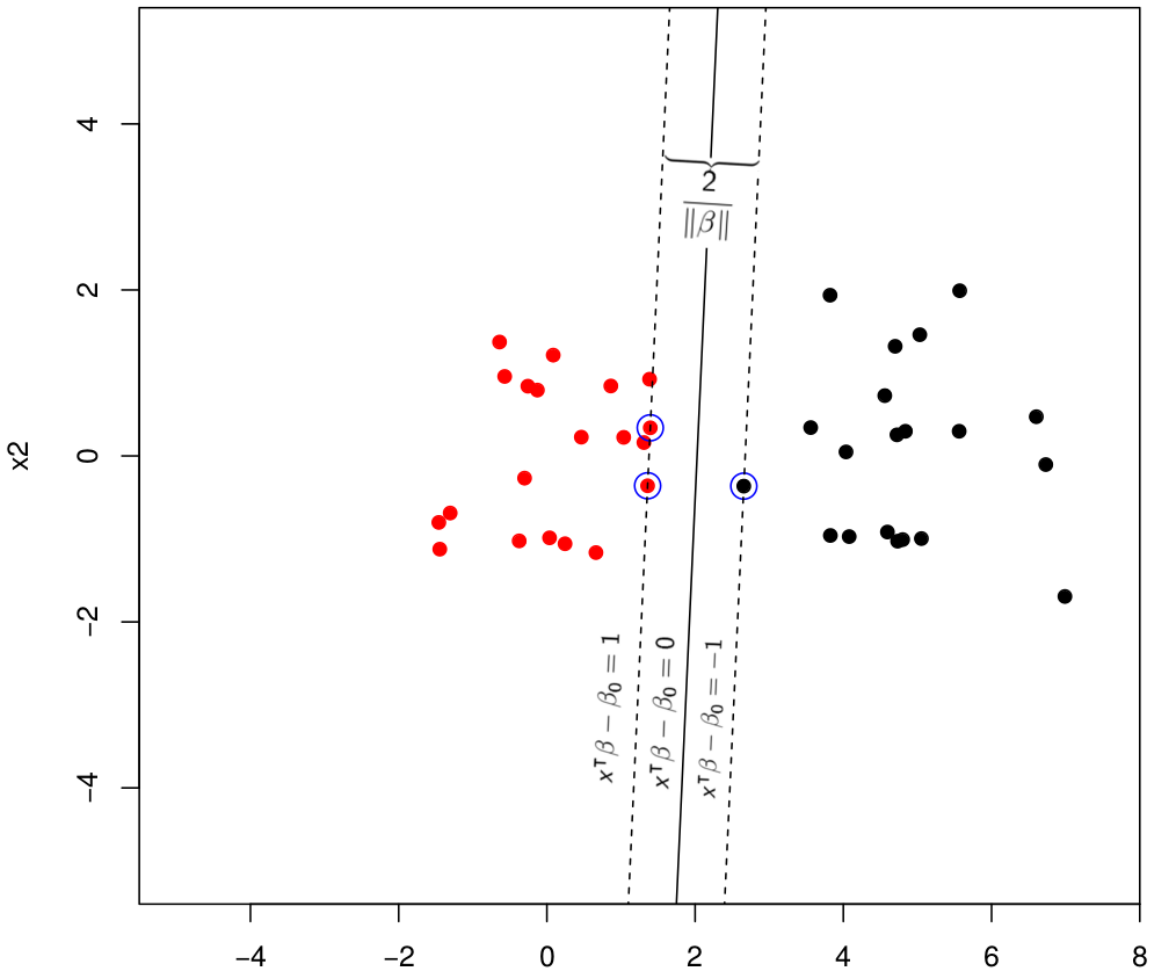
$$g(x) = x^\intercal \beta - \beta_0$$

можно построить классифицирующее правило

$$h(x) = \text{sign}\left[g(x)\right]$$

Из интуитивных соображений разумно полагать, что наилучшая разделяющая гиперплоскость – та, которая расположена дальше всего от представителей каждого из классов.

# Hard-margin SVM: margin, decision boundary



# Hard-margin SVM

Предположим, что данные – разделимы гиперплоскостью $x^\intercal \beta - \beta_0 = 0; \beta \in \mathbb{R}^p, \beta \in \mathbb{R}$.

Определив расстояние со знаком $g(x) = x^\intercal \beta - \beta_0$, можно считать, что $h(x_i) = \text{sign}\left[g(x)\right]$.

Из интуитивных соображений разумно полагать, что наилучшая разделяющая гиперплоскость – та, которая расположена дальше всего от представителей каждого из классов.

$\frac{2}{\|\beta\|}$ — растояние между парой гиперплоскостей, симметричных относительно разделяющей, таких, что никакая из точек не лежит между ними — называется отступом (margin).

$$\begin{cases} \frac{1}{2}\|\beta\|_2^2 \to \min\limits_{\beta, \beta_0} \\ y_i\left(x_i^\intercal \beta - \beta_0\right) \geq 1 \end{cases}$$

# Hard-margin SVM: множители Лагранжа

Воспользуемся методом множителей Лагранжа

$$\begin{cases} \inf\limits_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 - \sum\limits_{i=1}^{n} \alpha_i \left[ y_i \left( x_i^\intercal \beta + \beta_0 \right) - 1 \right] \to \max\limits_{\alpha_i} \\ \alpha_i \geq 0, \forall i \\ y_i \left( x_i^\intercal \beta - \beta_0 \right) \geq 1 \end{cases}$$

Так как оптимизируемая функция гладкая, можно воспользоваться необходимыми условиями экстремума

$$\frac{\partial}{\partial \beta} : \quad \beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial}{\partial \beta_0} : \quad 0 = \sum_{i=1}^{n} \alpha_i y_i$$

# Hard-margin SVM: двойственная задача Вольфа

Двойственная задача Вольфа:

$$\begin{cases} \frac{1}{2}\|\beta\|_2^2 - \sum_{i=1}^{n} \alpha_i \left[ y_i \left( x_i^\intercal \beta + \beta_0 \right) - 1 \right] \to \max\limits_{\alpha_i} \\ \alpha_i \geq 0, \forall i \\ \beta = \sum_{i=1}^{n} \alpha_i y_i x_i \\ 0 = \sum_{i=1}^{n} \alpha_i y_i \\ y_i \left( x_i^\intercal \beta - \beta_0 \right) \geq 1 \end{cases}$$

# Hard-margin SVM: двойственная задача Вольфа

Двойственная задача Вольфа:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\intercal x_j \to \max\limits_{\alpha_i} \\ \alpha_i \geq 0, \forall i \\ y_i \left( x_i^\intercal \beta - \beta_0 \right) \geq 1 \\ \beta = \sum_{i=1}^{n} \alpha_i y_i x_i \end{cases}$$

# Опорные вектора

В точке оптимума выполнены условия Каруша-Куна-Такера, в частности:

$$\alpha_i \left[1 - y_i \left(x_i^\intercal \beta - \beta_0\right)\right] = 0 \, \forall i$$

Т.е. либо

- $\alpha_i = 0 \Rightarrow y_i g\left(x_i\right) > 1$ – т.е. наблюдение не влияет на $\beta, \beta_0$
- $\alpha_i > 0 \Rightarrow y_i g\left(x_i\right) = 1$ – такое наблюдение будем называть опорным вектором

Из этих же соображений можно вычислить $\beta_0$ воспользовавшись произвольным опорным вектором.

# Slack variables

Обобщим процедуру на линейно-неразделимые данные
Позволим каждому из ограничений быть нарушеным на $\xi_i$:

$$y_i \left(x_i^\intercal \beta - \beta_0\right) \geq 1 - \xi_i, \xi_i \geq 0$$

ограничив при этом суммарную ошибку некоторым параметром

$$\sum_{i=1}^{n} \xi_i \leq t$$

Можно отметить, что

- $\xi_i = 0$ – наблюдения, лежащие вне зазора между классами
- $\xi_i \in (0; 1)$ – правильно классифицированные наблюдения, лежащие внутри зазора
- $\xi_i \geq 1$ – некорректно классифицированное наблюдения

# Прямая задача

Сформулируем оптимизационную задачу аналогично линейно-разделимому случаю:

$$\begin{cases} \frac{1}{2}\|\beta\|_2^2 \to \min\limits_{\beta,\beta_0} \\ y_i\left(x_i^\intercal \beta - \beta_0\right) \geq 1 - \xi_i \\ \sum\limits_{i=1}^{n} \xi_i \leq t \\ \xi_i \geq 0 \end{cases}$$

Величина отступа (margin), равная $\frac{2}{\|\beta\|}$ в данном случае определяет расстояние между парой гиперплоскостей, параллельных разделяющей, внутри которой присутствует пенальти за корректную классификацию.

Применяя те же соображения (использование множителей Лагранжа, двойственной задачи в форме Вольфа, условия Каруша-Куна-Такера), переходим к двойственной задаче.

# Двойственная задача

После необходимых преобразований, получается двойственная задача:

$$\begin{cases} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \to \max\limits_{\alpha_i} \\ \alpha_i \in [0; t] \\ \text{KKT+Wolfe conditions} \end{cases}$$

При этом, так же как и в линейно-разделимом случае

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

где $\alpha_i = 0$ для части наблюдений (в которых ограничение соблюдается строго)

# Влияние параметра $t$

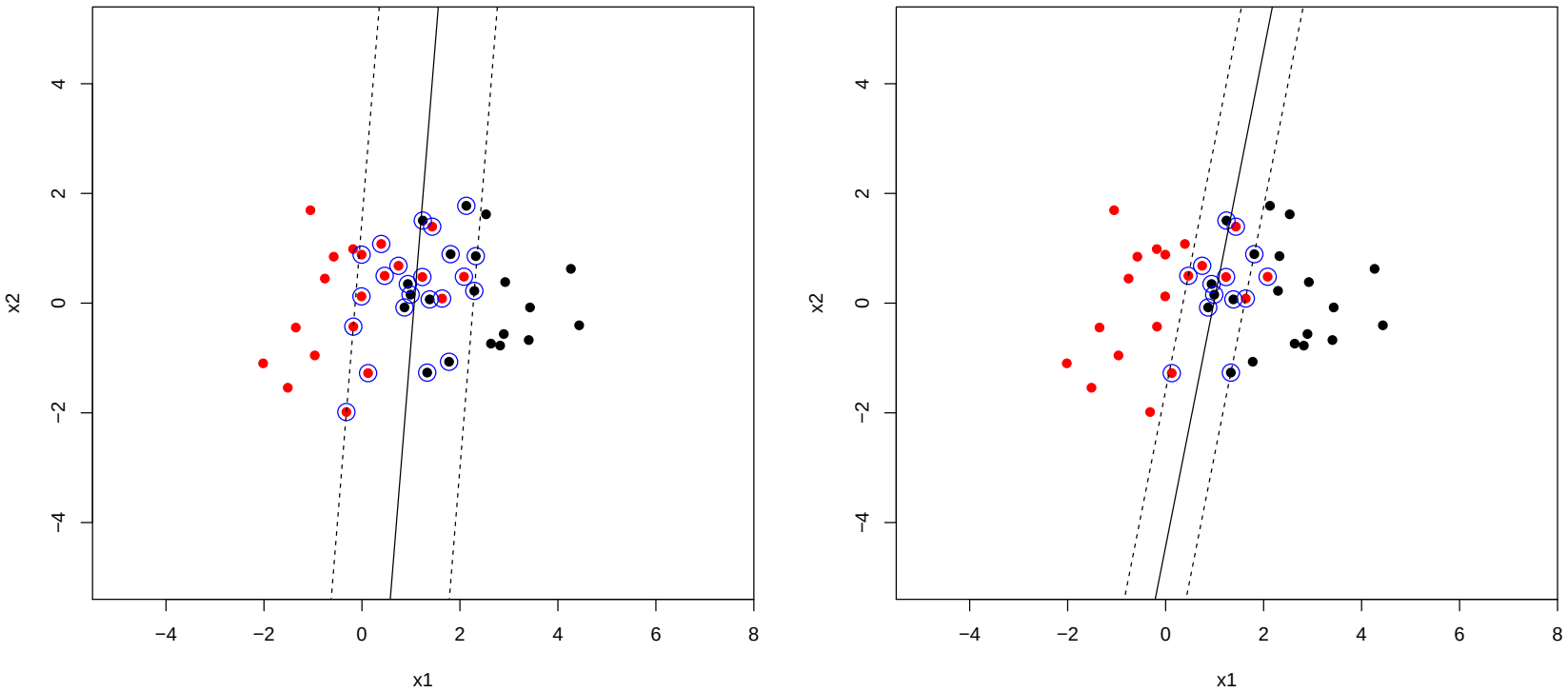



Рис.: Влияние максимально допустимой ошибки на SVM

# Эквивалентная переформулировка

Можно показать, что оптимизационная задача для soft-margin SVM эквивалентна задаче

$$\sum_{i=1}^{n} \max\left\{0, 1 - y_i\left(x_i^\intercal \beta - \beta_0\right)\right\} + \eta\|\beta\|_2^2 \to \min_{\beta, \beta_0}$$

Таким образом, SVM можно рассматривать как минимизацию эмпирического риска (с регуляризацией $\eta\|\beta\|_2^2$) и функцией потерь

$$\max\left\{0, 1 - y\left(x^\intercal \beta - \beta_0\right)\right\}$$

Которую можно рассматривать, как непрерывную аппроксимацию разрывной ошибки классификации

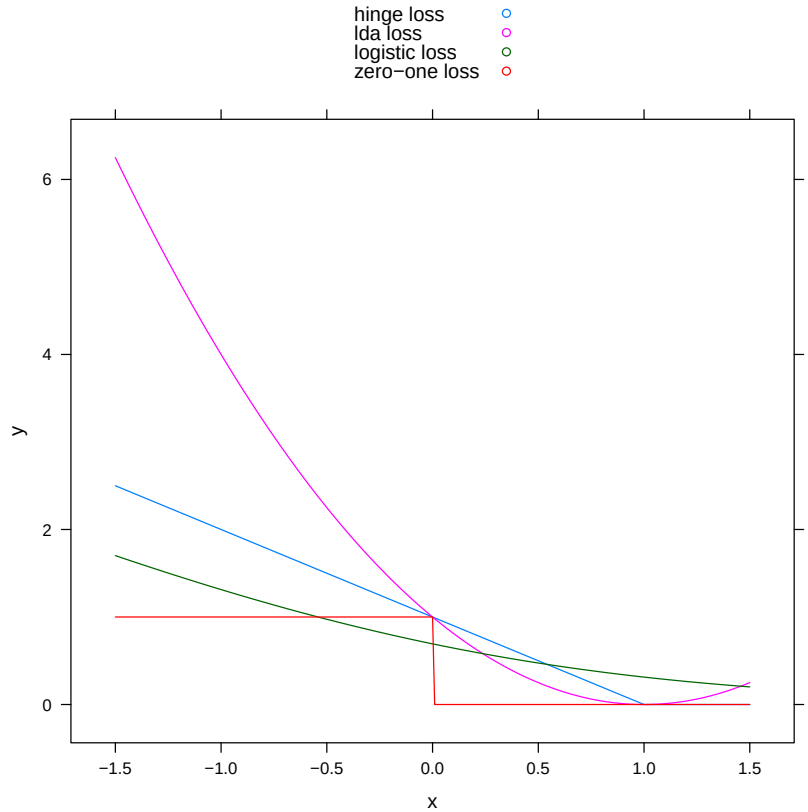
$$\mathbf{1}_{\left(h(x) \neq y\right)}$$

# Функция потерь в подходе Empirical Risk Minimization

Функции потерь:

- Логистическая регрессия $\log 1 + e^{-g(x)y}$
- SVM $\max\left\{0, 1 - g\left(x\right) y\right\}$
- Ошибка классификации: $\mathbf{1}_{y \neq \text{sign}\, g(x)}$
- LDA[a]: $\left(1 - g\left(x\right)\right)^2$

[a] В предположении одинаковых оценок $\hat{\pi}_i$



# Методы поиска решений

Для soft-margin SVM нами были получены две эквивалентные задачи:

- Прямая задача ($d+1$ параметров, $O(n)$ ограничений):

$$\begin{cases} \frac{1}{2}\|\beta\|_2^2 \to \min\limits_{\beta,\beta_0} \\ O(n) \text{ ограничений неравенства} \end{cases}$$

- Двойственная задача ($n$ параметров, $O(n)$ ограничений):

$$\begin{cases} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} \alpha_i\alpha_j y_i y_j x_i^\intercal x_j \to \max\limits_{\alpha_i} \\ O(n) \text{ ограничений неравенства} \end{cases}$$

В зависимости от соотношения $d$ и $n$ разумно использовать прямую или двойственную задачу.

# Поиск решений прямой задачи

Пользуясь эквивалентной формулировкой прямой задачи

$$\sum_{i=1}^{n} \max\left\{0, 1 - y_i\left(x_i^\intercal \beta - \beta_0\right)\right\} + \eta\|\beta\|_2^2 \to \min_{\beta,\beta_0}$$

Предлагается производить оптимизацию по $\beta$ используя стохастический градиентный спуск с аккуратным выбором величины шага.

Можно показать[1], что такой метод позволяет получить решение с точностью $\varepsilon$ за $O\left(\frac{1}{\varepsilon}\right)$ итераций; при этом количество наблюдений не влияет на асимптотику количества итераций (но сложность одной итерации зависит от количества наблюдений).

[1] Pegasos: primal estimated sub-gradient solver for SVM

# Поиск решений двойственной задачи

Для двойственной задачи

$$\mathcal{F}(\alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\intercal x_j \to \max_{\alpha_i}$$

Предлагается[2] совершить несколько итераций покоординатного градиентного спуска:

- Каждый $\alpha_i$ изменяется в направлении $\frac{\partial \mathcal{F}}{\partial \alpha_i}$
- Очередное решение проецируется на множество допустимых
- Поддерживается необходимая для вычисления $x^\intercal \beta$ информация

Последовательность решений сходится как минимум линейно

[2] A Dual Coordinate Descent Method for Large-scale Linear SVM

# Спрямляющее пространство

По построению, SVM применим лишь к (почти) линейно-разделимым данным.

Предположим, что известен набор отображений

$$\left\{\varphi_j(x)\right\}_{j=1}^{d}$$

$$\varphi_j : \mathbb{R}^p \to \mathbb{R}$$

таких, что набор данных

$$\left\{\tilde{x}_i = \begin{pmatrix} \varphi_1(x_i) \\ \vdots \\ \varphi_d(x_i) \end{pmatrix}, y_i\right\}_{i=1}^{n}$$

линейно разделим; будем называть пространство, в котором наблюдается разделимость, спрямляющим пространством

# Двойственная задача в спрямляющем пространстве

Записывая двойственную задачу в спрямляющем пространстве,

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \tilde{x}_i^T \tilde{x}_j \to \max_{\alpha_i}$$

можно заметить, что $\tilde{x}_i$ входят в её формулировку лишь в виде скалярных произведений $\tilde{x}_i^T \tilde{x}_j$.

Также, используя выражение

$$\beta = \sum_{i=1}^{n} \alpha_i y_i \tilde{x}_i$$

вычисление расстояния до разделяющей гиперплоскости (со знаком) также сводится к использованию скалярных произведений $\tilde{x}_i$.

# Kernel-trick: Теорема Мерсера

Таким образом, скалярного произведения в спрямляющем пространстве достаточно для построения (и применения) классифицирующего правила.

Кроме того, знание отображений $\varphi_j$ не требуется и спрямляющее пространство может быть бесконечномерным.

# Kernel-trick: Теорема Мерсера

**Теорема**

*Пусть* $K(u,v): X \times X \to \mathbb{R}$ *– отображение:*

- *Симметричное:* $K(u,v) = K(v,u)$
- *Положительно определённое:* $\iint\limits_{X \times X} K(u,v)\, g(u)\, g(v)\, \mathrm{d}u\, \mathrm{d}v \geq 0, \forall g : X \to \mathbb{R}$

*Тогда (и только тогда) существует пространство* $H$ *и отображение* $\varphi : X \to H : K(u,v) = \langle \varphi(u), \varphi(v) \rangle_H$

Таким образом, для любой функции, являющейся ядром, существует пространство со скалярным произведением.

Цель использования данного соображения применительно к SVM в том, что после перехода в новое пространство (с помощью $\varphi$) исходные данные могут стать почти линейно-разделимыми.

# Kernel-trick: операции над ядрами

Известны некоторые способы конструирования ядер, позволяющие не проверять условия теоремы Мерсера:

- Скалярное произведение в векторном пространстве
- Положительная константа
- Произведение ядер: $K(u,v) = K_1(u,v) K_2(u,v)$
- Произведение отображений: $K(u,v) = \varphi(u)\varphi(v), \varphi : X \to \mathbb{R}$
- Линейная комбинация с положительными коэффициентами: $K(u,v) = \alpha_1 K_1(u,v) + \alpha_2 K_2(u,v), \alpha_{1,2} > 0$
- Композиция ядра и отображения: $K(u,v) = K_1\left(\varphi(u), \varphi(v)\right)$
- Степенной ряд:

$$f : \mathbb{R} \to \mathbb{R}$$

сходящийся степенной ряд с положительными коэффициентами, тогда

$$K(u,v) = f\left(K_1(u,v)\right)$$

является ядром

# Kernel-trick: примеры ядер

Используя предыдущие рассуждения, несложно показать, что ядрами являются:

- Полиномиальное ядро: $K(u, v) = \left(1 + \langle u, v \rangle\right)^d$ (базисные функции – мономы степени $\leq d$; размерность пространства – $C_{p+d}^d$
- Ядро радиальных базисных функций: $K(u, v) = e^{-\frac{\|u-v\|_2^2}{2\sigma^2}}$

Выбор ядра может быть обусловлен либо знанием о данных (например эмпирическим, в случае если они похожи на отделимые поверхностями соответствующего порядка), либо производится автоматически на основе кросс-валидации или иной процедуры

# Множественная классификация: one vs one

Построение SVM происходило для классификации с двумя классами; покажем, как можно обобщить полученное решение на классификацию с $N$ классами:

- Для всех $\frac{N(N+1)}{2}$ пар классов построим SVM-классификатор
- Обозначим за $N_i(x)$ количество парных классификаций, в которых был выбран класс $i$

Классифицирующее правило: $g(x) = \underset{i}{\operatorname{argmax}} N_i(x)$

# Множественная классификация: one vs other

Покажем иное построение классификации с $N$ классами, использующее меньшее количество классификаторов.

- Построим $N$ классификаторов для задач классификации

$$y_i = 1 \Leftrightarrow x_i \text{ из } i\text{-го класса}$$

- Соответствующее классифицирующее правило обозначим

$$g_i(x) = \text{sign}\left(h_i(x)\right)$$

Классифицирующее правило: $g(x) = \underset{i}{\text{argmax}}\, h_i(x)$

# Регрессия с помощью SVM

Сформулируем задачу регрессии в виде, похожем на SVM:

$$\begin{cases} C\sum_{i=1}^{n}\left(\xi_i^+ + \xi_i^-\right) + \frac{1}{2}\|\beta\|_2^2 \to \min\limits_{\beta,\beta_0} \\ \xi_i^+, \xi_i^- \geq 0 \\ y_i - \left(x_i^\intercal \beta + \beta_0\right) \leq \varepsilon + \xi_i^+ \\ -y_i + \left(x_i^\intercal \beta + \beta_0\right) \leq \varepsilon + \xi_i^- \end{cases}$$

Оказывается, что в такой формулировке

- $\beta$ описывается как линейная комбинация опорных векторов
- В двойственной задаче $x_i$ встречаются только в виде скалярных произведений $\Rightarrow$ можно использовать kernel-trick

# Изменение регуляризации в SVM

С целью отбора признаков можно изменить регуляризацию в одной из эквивалентных формулировок SVM:

$$C\sum_{i=1}^{n}\max\left\{0, 1 - y_i\left(x_i^\intercal\beta - \beta_0\right)\right\} + \Phi\left(\beta\right) \to \min_{\beta,\beta_0}$$

- LASSO SVM: $\Phi\left(\beta\right) = \|\beta\|_1$
- Doubly-regularized SVM: $\Phi\left(\beta\right) = \alpha\|\beta\|_1 + \frac{1}{2}\|\beta\|_2^2$
- Support Features Machine: $\Phi\left(\beta\right) = \sum_{i=1}^{p}\max\left\{2\mu\beta_i, \mu^2 + \beta_i^2\right\}$
- Relevance Features Machine: $\Phi\left(\beta\right) = \sum_{i=1}^{p}\ln\left(\beta_i^2 + \frac{1}{\mu}\right)$

($\alpha, \mu$ — дополнительные параметры; коэффициент soft-margin SVM $t$, эквивалентным преобразованием перемещён к $\sum$)

# Выбор параметров с помощью кросс-валидации

SVM требует выбора максимально возможного нарушения ограничений $t$, в случае использования kernel-trick — выбора ядра и, возможно, параметров ядра; а также — параметров регуляризатора в SVM-подобных процедурах.

Для выбора оптимального набора параметров предлагается воспользоваться кросс-валидацией:

- Данные разделяются на $K$ частей сходного размера
- Для всех $k = 1, \ldots, K$ строится оценка $\hat{\beta}^{(k)}, \hat{\beta}_0^{(k)}$ по данным из всех частей, кроме $k$-ой
- Для всех $k$ по отброшенной части данных строится оценка ошибочной классификации: $\varepsilon_k = \frac{n_k}{n} \mathbf{1}_{y_i \neq y_i^{(k)}}$
- Строится оценка качества классификации $\mathcal{E} = \sum_k \varepsilon_k$

Из всего рассматриваемого пространства параметров выбирается набор параметров, минимизирующих $\mathcal{E}$; после чего с использованием этого набора параметров строится модель по всем данным.

# Некоторые соображения о кросс-валидации

Терминология:

- K-fold cross-validation – кросс-валидация с разделением на $K$ частей
- Leave-one-out cross-validation – кросс-валидация с разделением на одноэлементные множества

Полезные соображения:

- Соображения кросс-валидации применимы и к регрессии, в этом случае может использоваться оценка

$$\varepsilon_k = \frac{n_k}{n}\mathrm{MSE}_k$$

- Важно применять кросс-валидацию ко всей процедуре оценки параметров в целом, не допуская использования всех доступных данных на каком-либо промежуточном этапе
- Оценка качества классификации (MSE регрессии) получается смещённой (с положительным смещением), так как оценка параметров производится по меньшим наборам данных

В рамках SVM можно предъявить оценку, связывающую математическое ожидание ошибки классификации и эмпирическую ошибку классификации.

Пусть $x_i, y$ – выборка из распределения $P(x, y)$, определим:

- $y = f(x, \alpha)$ – модель классификатора, зависящая от $\alpha$
- Математическое ожидание ошибки классификации

$$R(\alpha) = \int \mathbf{1}_{y \neq f(x,\alpha)} \, \mathrm{d}P(x, y)$$

- Эмпирическая ошибка классификации

$$R_e(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{y_i \neq f(x_i,\alpha)}$$

При этом с вероятностью $1 - \eta$, $0 < \eta < 1$ выполнено:

$$R(\alpha) \leq R_e(\alpha) + \sqrt{\frac{h\left(1 + \log\frac{2n}{h}\right) - \log\frac{\eta}{4}}{n}}$$

Где $h$ – размерность Вапника-Червоненкиса (VC-размерность), характеризующая сложность семейства алгоритмов для классификации с двумя классами.

- Для гиперплоскостей в $\mathbb{R}^p$: $h = p + 1$
- Если любой набор из $k$ точек разделим для любого назначения классов $\Rightarrow h \geq k$
- С помощью данного факта можно получить оценку сверху на математическое ожидание ошибки классификации для некоторых классов функций
- Легко видеть, что при $n \to \infty$ оценка сходится к $R_e$

# Сравнение SVM с LDA

Оптимизационная задача SVM:

$$\begin{cases} \frac{1}{2}\|\beta\|_2^2 \to \min\limits_{\beta,\beta_0} \\ y_i\left(x_i^\intercal \beta - \beta_0\right) \geq 1 - \xi_i \\ \sum\limits_{i=1}^{n} \xi_i \leq t \\ \xi_i \geq 0 \end{cases}$$

Оптимизационная задача LDA:

$$\begin{cases} \beta^\intercal \Sigma_B \beta \to \max \\ \beta^\intercal \Sigma_W \beta = 1 \end{cases}$$

- LDA (в построении) предполагает нормальность; SVM — свободен от предположений о распределении данных
- На построение решающего правила при помощи LDA влияют все наблюдения, при использовании SVM – только опорные вектора
- В LDA (в отличие от SVM) априрорные вероятности принадлежности классу влияют на сдвиг границы классификации
- LDA имеет аналитическое решение (с помощью обобщённых собственных векторов)
- Оба метода допускают использование kernel-trick