Support Vector Machines

Д. Корчемкин, В. Агеев
622 группа

26 ноября 2017 г.

1 SVM

Будем рассматривать задачу классификации в рамках обучения с учителем.

Имеется выборка $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$; задачей является построение классифицирующего правила $f : \mathbb{R}^p \to \{-1, 1\}$.

Отметим, что никаких дополнительных ограничений на распределение не требуется.

1.1 Hard-margin SVM

Предположим, что присутсвует линейная разделимость, т.е. существует гиперплоскость (определяемая уравнением $x^\intercal \beta - \beta_0 = 0$ ($x, \beta \in \mathbb{R}^p; \beta_0 \in \mathbb{R}$), такая, что точки, соответствующие разным классам лежат в различных полупространствах относительно гиперплоскости.

Факт принадлежности наблюдений из разных классов разным полупространствам можно (возможно, изменив знаки β, β_0) описать уравнениями:

$$\begin{bmatrix} x_i^\intercal \beta - \beta_0 < 0 & y_i = -1 \\ x_i^\intercal \beta - \beta_0 > 0 & y_i = 1 \end{bmatrix} \Leftrightarrow \left(x_i^\intercal \beta - \beta_0\right) y_i > 0$$

В таком случае, классифицирующим правилом разумно принять

$$g(x) = \operatorname{sign}\left(x^\intercal \beta - \beta_0\right)$$

Ясно, что в случае линейно разделимых данных может существовать более одной гиперплоскости, разделяющей данные. Введём критерий оптимальности: максимальное расстояние между двумя гиперплоскостями, параллельных данной и симметрично расположенных относительно неё, при котором между ними не находится ни одна из точек x_i; это расстояние будем называть зазором (margin).

Легко видеть, что каждой из двух параллельных гиперплоскостей будет принадлежать некоторое количество точек из соответствующего класса (иначе, так как количество точек в выборке конечно, то растояние между гиперплоскостями можно

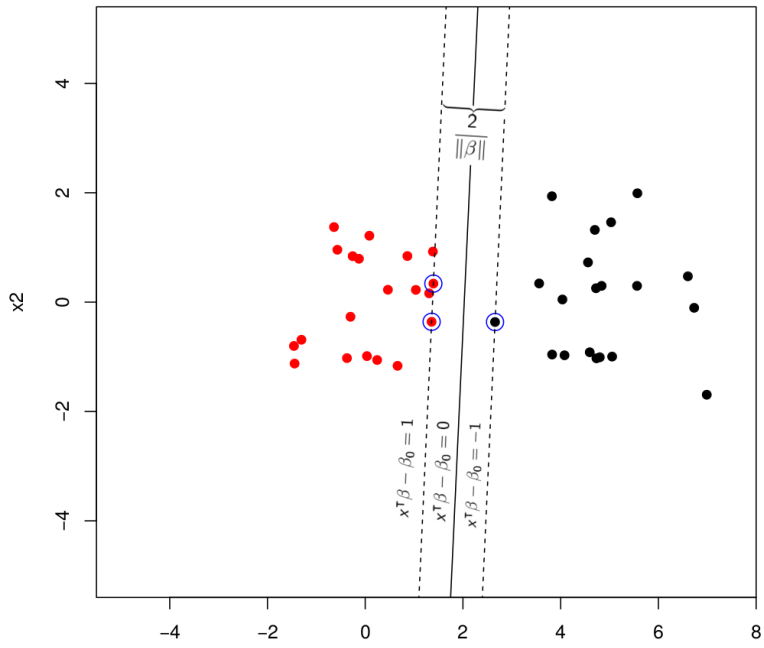

Рис. 1: Hard-margin SVM

увеличить, сместив гиперплоскость, которой не принадлежит ни одной точки); точки, которые принадлежат одной из гиперплоскостей — будем называть опорными векторами.

С точностью до нормировки вектора β эта пара гиперплоскостей может быть описана парой уравнений:

$$x^\intercal\beta - \beta_0 = -1$$
$$x^\intercal\beta - \beta_0 = 1$$

а растояние между ними составит $\frac{2}{\|\beta\|}$ (см. поясняющий рисунок 1).

Принадлежность точек обучающей выборки полупространствам описывается уравнениями

$$\begin{bmatrix} x_i^\intercal\beta - \beta_0 \leq 1 & y_i = -1 \\ x_i^\intercal\beta - \beta_0 \geq 1 & y_i = 1 \end{bmatrix} \Leftrightarrow \left(x_i^\intercal\beta - \beta_0\right) y_i \geq 1$$

Таким образом, в случае линейно разделимой выборки, задача выбора оптимальной гиперплоскости сводится к следующей задаче квадратичного программирования с линейными ограничениями:

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

Пользуясь принципом Лагранжа, из (1) получаем задачу:

$$\begin{cases} \inf\limits_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 - \sum\limits_{i=1}^{n} \alpha_i \left(y_i \left(x_i^\intercal \beta - \beta_0 \right) - 1 \right) \to \max\limits_{\alpha_1,\ldots,\alpha_n} \\ \alpha_i \geq 0, \forall i \end{cases} \tag{2}$$

Так как все функции гладкие, то inf достигается в точке, в которой выполнены необходимые условия экстремума:

$$\begin{aligned} \tfrac{\partial}{\partial \beta} : &\quad \beta = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \tfrac{\partial}{\partial \beta_0} : &\quad 0 = \sum_{i=1}^{n} \alpha_i y_i \end{aligned}$$

подставляя эти равенства в (2) получаем:

$$\begin{cases} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{k=1}^{n} \alpha_i \alpha_k y_i y_k x_i^\intercal x_k \to \max\limits_{\alpha_1,\ldots,\alpha_n} \\ \alpha_i \geq 0, \forall i \\ \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i \left(y_i \left(x_i^\intercal \beta - \beta_0 \right) - 1 \right) = 0 \end{cases} \tag{3}$$

Из условия регулярности ККТ:

$$\alpha_i \left(y_i \left(x_i^\intercal \beta - \beta_0 \right) - 1 \right) = 0 \tag{4}$$

в оптимальной точке, т.е. либо $\alpha_i = 0$, либо x_i является опорным вектором (принадлежит одной из пары плоскостей, описаных выше).

Таким образом, решающее правило строится на основе «сложных» для классификации наблюдений, а остальные наблюдения — не влияют (явным образом) на расположение разделяющей гиперплоскости.

1.2 Soft-margin SVM

Понятно, что требование линейной разделимости классов слишком сильное для реальной применимости SVM как метода классификации.

Позволим для этого каждому из ограничений в задаче (1) быть несколько ослабленным:

$$\begin{cases} \frac{1}{2}\|\beta\|_2^2 \to \min\limits_{\beta,\beta_0} \\ \left(x_i^\intercal \beta - \beta_0 \right) y_i \geq 1 - \xi_i, \forall i \\ \sum\limits_{i=1}^{n} \xi_i \leq t \\ \xi_i \geq 0 \, \forall i \end{cases} \tag{5}$$

(t — параметр алгоритма; линейно-разделимый случай соответствует $t = 0$).

Далее, повторяя все рассуждения для линейно-разделимого случая, используем принцип Лагранжа:

$$\begin{cases} \inf\limits_{\beta,\beta_0,\xi_1,\ldots,\xi_n} \frac{1}{2}\|\beta\|_2^2 + \lambda \sum\limits_{i=1}^{n} \xi_i - \sum\limits_{i=1}^{n} \alpha_i \left(y_i \left(x_i^\intercal \beta - \beta_0 \right) - (1 - \xi_i) \right) - \sum\limits_{i=1}^{n} \gamma_i \xi_i \to \max\limits_{\alpha_1,\ldots,\alpha_n,\gamma_1,\ldots,\gamma_n,\lambda} \\ 0 \le \alpha_i \le t, \forall i \\ \gamma_i \ge 0, \forall i \\ \lambda \ge 0 \end{cases} \tag{6}$$

Опять же, ввиду гладкости, inf достигается в точке, в которой выполнены необходимые условия экстремума:

$$\begin{aligned} \tfrac{\partial}{\partial \beta} : &\quad \beta = \sum_{i=1}^{n} \alpha_i y_i x_i \\ \tfrac{\partial}{\partial \beta_0} : &\quad 0 = \sum_{i=1}^{n} \alpha_i y_i \\ \tfrac{\partial}{\partial \xi_i} : &\quad \alpha_i = \lambda - \gamma_i \end{aligned}$$

используя эти равенства в (6), получаем:

$$\begin{cases} \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{k=1}^{n} \alpha_i \alpha_k y_i y_k x_i^\intercal x_k \to \max\limits_{\alpha_1,\ldots,\alpha_n} \\ 0 \le \alpha_i \le t, \forall i \\ \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{7}$$

Из условия регулярности ККТ опять же следует классификация векторов на опорные, аутлаеры и не участвующие в построении правила предсказания точки.

Параметр t, определяющий допустимое нарушение ограничений, позволяет варьировать количество опорных векторов (которые, вследствие ККТ-условий, соответствуют векторам, участвующим в построении классифицирующего правила; пример изменения разделяющей гиперплоскости и набора опорных векторов показан на рисунке 2).

1.3 Методы оптимизации

В зависимости от соотношения размерностей n и p, может быть выгодно (с точки зрения вычислительных затрат) решать прямую (5), или двойственную 7 задачу.

При этом, несмотря на то, что обе задачи являются задачами квадратичного программирования с линейными ограничениями неравенства, использование методов общего вида осложняется экспоненциальной (в худшем случае) сложностью.

Для формулировки прямой задачи в виде

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

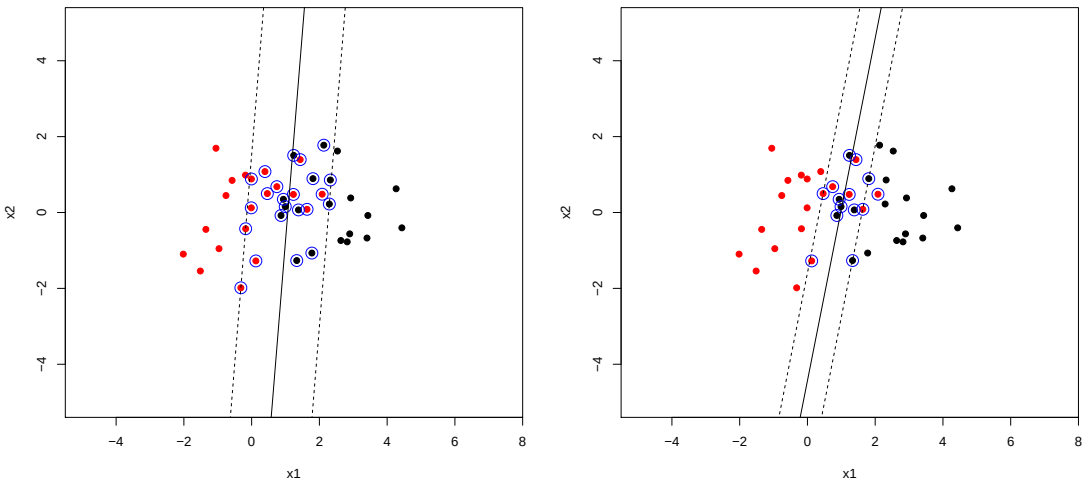

Рис. 2: Влияние максимально допустимой ошибки на soft-margin SVM

предлагается использовать (стохастический) градиентный спуск со специфическим (для задачи SVM) выбором величины шага (Pegasos: primal estimated sub-gradient solver for SVM); в этом случае можно показать, что решение с точностью ε достигается за $O\left(\frac{1}{\varepsilon}\right)$ итераций; при этом количество наблюдений не влияет на асимптотику количества итераций.

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 (A Dual Coordinate Descent Method for Large-scale Linear SVM) использовать покоординатный градиентный спуск:

- Каждый α_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

1.4 SVM как частный случай Empirical Risk Minimization

Можно доказать, что задача SVM эквивалентна задаче

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

где C – параметр алгоритма; домножив на $\frac{1}{nC}$ получается эквивалентная задача

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

В такой формулировке можно сравнить SVM, логистическую регрессию и LDA как непрерывные (SVM) и гладкие (LDA, логистическая регрессия) аппроксимации ошибки классификации (в смысле методов минимизации [регуляризованного] эмпирического риска) следующими функциями потерь (графики функций потерь приведены на рисунке 3):

- SVM (hinge loss): $\max\left\{1-\left(x^{\intercal}\beta+\beta_0\right)y_i,0\right\}$
- LDA[1]: $\left(1-y_i\left(\beta_0+\beta^{\intercal}x_i\right)\right)^2$
- Logistic regression: $\log\left(1+e^{-y_i(\beta^{\intercal}x_i+\beta_0)}\right)$

Можно заметить, что LDA и логистическая регрессия имеют штраф и за правильно классифицированные наблюдения; для логистической регрессии этот штраф убывает при удалении от разделяющей гиперплоскости, а для LDA — начинает увеличиваться при удалении от центра класса.

2 Расширения SVM

2.1 Kernel trick

Описаный выше алгоритм может быть использован лишь в случае, когда данные относительно похожи на линейно-разделимые.

Предположим, что существует некоторое отображение $\Phi:\mathbb{R}^p\to V$ где V – некоторое гильбертово пространство.

Тогда, несложно заметить, что применяя SVM к образам исходных векторов, мы получаем задачу квадратичного программирования:

$$\begin{cases}\sum\limits_{i=1}^{n}\alpha_i-\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{n}\alpha_i\alpha_k y_i y_k\left\langle\Phi\left(x_i\right),\Phi\left(x_k\right)\right\rangle_V\to\max\limits_{\alpha_1,\ldots,\alpha_n}\\ 0\le\alpha_i\le t,\forall i\\ \sum\limits_{i=1}^{n}\alpha_i y_i=0\end{cases}$$

и классифицирующая функция

$$h\left(x\right)=\operatorname{sign}\left[\sum_{i=1}^{n}\alpha_i y_i\left\langle\Phi\left(x_i\right),\Phi\left(x\right)\right\rangle_V-\beta_0\right]$$

[1]В случае $\sum y_i=0$ можно показать следующую цепочку эквивалентных переходов: LDA $\Leftrightarrow$ FDA $\Leftrightarrow$ CCA (с 1 переменной в одном из наборов признаков) $\Leftrightarrow$ OLS $\sum\left(y_i-\left(\beta_0+\beta^{\intercal}x_i\right)\right)^2\to\min_{\beta,\beta_0}\Leftrightarrow$ *ERM*

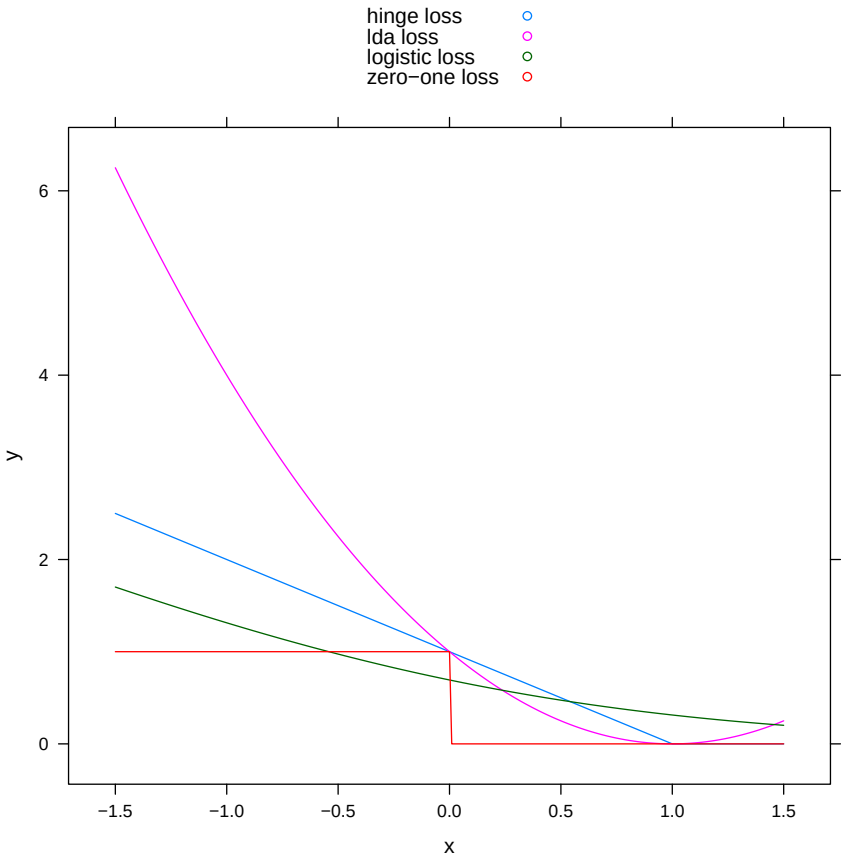

Рис. 3: Различные функции потерь как непрерывные аппроксимации ошибки классификации

(правило классификации остаётся прежним — $\operatorname{sign}\left(h\left(x\right)\right)$

Можно заметить, что во всех выражениях результат применения Φ используется только для использования в скалярном произведении с результатом применения Φ к другому вектору из $\mathbb{R}^p$, что позволяет (используя теорему Мерсера) использовать произвольную симметричную положительно определённую функцию $k\left(u,v\right) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ (ядро) в качестве скалярного произведения в некотором векторном пространстве; в этом случае преобразование Φ может соответствовать отображению, состоящему из собственных функций k:

$$\begin{aligned} k\left(u,v\right) &= \sum_{i=1}^{\infty} \theta_j \varphi_k\left(u\right) \varphi_k\left(v\right) \\ \Phi\left(x\right) &= \left[\varphi_1\left(x\right), \ldots, \varphi_k\left(x\right), \ldots, \right] \end{aligned} \tag{8}$$

Данное соображение позволяет применять SVM к данным в достаточной степени линейно разделимым в некотором гильбертовом пространстве (в том числе — бесконечномерном); в том числе — без предъявления в явном виде отображения из исходного пространства в данное.

Так как непосредственная проверка положительной определённости ядра представляет собой проблему, можно воспользоваться следующими операциями, приводящими к получению новых ядер:

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

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

  сходящийся степенной ряд с положительными коэффициентами, тогда

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

  является ядром

Часто используемые ядра:

- RBF (radial basis functions): $k\left(u,v\right) = e^{-\gamma\|u-v\|_2^2}$
- Полиномиальное (степеней $\leq d$: $k\left(u,v\right) = \left(\langle u,v\rangle + 1\right)^d$

Используя (8) можно показать, что использование полиномиального ядра степени d соответствует отображению в C_{p+d}^d-мерное пространство, гиперплоскостям в котором будут соответствовать поверхности порядка d в исходном пространстве.

Рассматривая RBF-ядро, можно убедиться, что $k(u,v) \xrightarrow[\gamma\to\infty]{} \mathbf{1}_{u=v}$; что позволяет добиться линейной разделимости для произвольного набора данных (предполагая отсутствие наблюдений с одинаковыми значениями признаков, принадлежащим разным классам).

2.2 Изменение регуляризации

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

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

- LASSO SVM: $\Phi(\beta)=\|\beta\|_1$
  - Чем меньше C, тем больше влияние ℓ_1-регуляризации
  - Может попеременно отбрасывать и шумовые и значимые признаки при варьировании C
  - Зависимые признаки не группируются
- Doubly-regularized SVM: $\Phi(\beta)=\alpha\|\beta\|_1+\frac{1}{2}\|\beta\|_2^2$
  - Чем больше α, тем больше влияние ℓ_1-регуляризации
  - Присутствует эффект группировки
  - Может попеременно отбрасывать и шумовые и значимые признаки при варьировании C, α
- Support Features Machine: $\Phi(\beta)=\sum\limits_{i=1}^{p}\max\left\{2\mu\beta_i, \mu^2+\beta_i^2\right\}$
  - μ – параметр «селективности»
  - Присутствует эффект группировки
  - Значимые ($\left\|\beta_j\right\|>\mu$) признаки группируются и входят в решение совместно
  - Шумовые признаки ($\left\|\beta_j\right\|<\mu$) подавляются
- Relevance Features Machine: $\Phi(\beta)=\sum\limits_{i=1}^{p}\ln\left(\beta_i^2+\frac{1}{\mu}\right)$
  - μ – параметр «селективности»
  - Присутствует эффект группировки

– Лучше выбирает набор значимых признаков

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

2.3 Support Vector Regression

В секциях 1.1-1.2 было показано, что классифицирующее правило в SVM зависит только от «сложно»-классифицируемых наблюдений (опорных векторов).

Оказывается, такую же идею можно использовать и для задач регрессии:

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

(где ε – параметр алгоритма); добавляя возможность нарушения ограничений, переходим к задаче

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

Можно показать, что двойственной к ней является задача

$$\begin{cases} \frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\left(\alpha_i^+ - \alpha_i^-\right)\left(\alpha_j^+ - \alpha_j^-\right)x_i^\intercal x_j + \varepsilon\sum\limits_{i=1}^{n}\left(\alpha_i^+ + \alpha_i^-\right) + \sum\limits_{i=1}^{n} y_i\left(\alpha_i^- - \alpha_i^+\right) \to \min \\ \sum\limits_{i=1}^{n}\left(\alpha_i^+ - \alpha_i^-\right) = 0 \\ 0 \le \alpha_i^+ \le C \\ 0 \le \alpha_i^- \le C \\ \beta = \sum\limits_{i=1}^{n}\left(\alpha_i^+ - \alpha_i^-\right)x_i \end{cases}$$

При этом, в силу ККТ-условий, многие $\alpha_i = 0$, т.е. в построении функции регрессии как и в SVM участвует лишь ограниченное количество опорных наблюдений.

Аналогично SVM, двойственная задача регрессии допускает нелинейное обобщение с использованием kernel trick.

2.4 Multi-class SVM

В предъявленом построении SVM рассматривался лишь случай классификации с двумя классами.

Для распространения идеи SVM на классификацию с большим количеством классов существует несколько подходов, в частности:

- Классификация с использованием сравнений вида “один со многими”
- Классификация с использованием сравнений вида “каждый с каждым”

2.4.1 Сравнения “один со многими”

Для классификации с N классами строится N классифицирующих правил $h_i(x)$; кодирующих принадлежность i-му классу за 1, а принадлежность любому другому классу за -1.

В качестве результирующего решающего правила используется

$$h(x) = \operatorname*{argmax}_i h_i(x)$$

2.4.2 Сравнения “каждый с каждым”

Для классификации с N классами строится $\frac{N(N-1)}{2}$ классифицирующих правил, производящих классификацию для каждой возможной пары классов.

Обозначив за N_i количество сравнений, в которых элемент x был классифицирован как принадлежащий i-ому классу; в качестве классифицирующего правила предлагается использовать

$$h(x) = \operatorname*{argmax}_i N_i$$

3 Выбор параметров

Ввиду наличия свободы выбора значения параметра регуляризации, ядра (или семейства ядер), необходимо предъявить процедуру сравнения построенных классификаторов. Так как в предлагаемой процедуре не используются никакие предположения о распределении $P(x, y)$, использование информационных критериев (BIC, AIC, ...) для этих целей невозможно.

Предлагается рассмотреть процедуры выбора параметров, основанные на (общей) идее кросс-валидации и специфичную для SVM оценку эмпирического риска на основе комбинаторной размерности.

3.1 Кросс-валидация

Предполагая, что выборка является выборкой из распределения $\langle x_i, y_i \rangle \sim \mathcal{P}(x, y)$, с точки зрения выбора параметров или типа алгоритма классификации (или регрессии), хотелось бы получать не только выборочную оценку качества классификации (регрессии), но и оценку пригодности выбранного алгоритма в смысле качества классификации (регрессии) на всём распределении $\mathcal{P}$, а не только конкретной выборки из него.

Для этого предлагается построить несколько моделей по различным подмножествам датасета, оценить для каждой из моделей качество классификации (регрессии) по части датасета, не участвовавшей в оценке параметров, после чего получить оценку математического ожидания качества классификации (регрессии).

Обозначив за $N = \{1, \ldots, n\}$ множество индексов элементов выборки, выберем K подмножеств (примерно одинакового размера) $N_k \subset N$.

Построим K классификаторов (функций регрессии) $f_k(x) = f\left(x, \hat{\theta}\left(N \setminus N_k\right)\right)$.

3.1.1 Классификация

Для каждой из них вычислим эмпирическую ошибку классификации

$$\hat{\varepsilon}_k = \frac{1}{\#N_k} \sum_{i \in N_k} \mathbf{1}_{f_k(x_i) \neq y_i}$$

Используя данную оценку, построим оценку ошибки классификации

$$\hat{\varepsilon} = \frac{1}{n} \sum_{i=1}^{K} \#N_k \hat{\varepsilon}_k$$

Также можно оценить «разброс» ошибки классификации как

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{1}{K-1} \sum_{i=1}^{K} \left(\hat{\varepsilon}_k - \hat{\varepsilon}\right)^2}$$

3.1.2 Регрессия

Для каждой из построенных моделей вычислим ошибку регрессии

$$\hat{\varepsilon}_k = \frac{1}{\#N_k} \sum_{i \in N_k} \left(f_k(x_i) - y_i\right)^2$$

И используя данные оценки, построим оценку дисперсии остатка регрессии для семейства f:

$$\hat{\varepsilon} = \frac{1}{n} \sum_{i=1}^{K} \#N_k \hat{\varepsilon}_k$$

3.1.3 Общие соображения

Получив оценки $\hat{\varepsilon}$ для всех интересующих параметрических семейств классификаторов (функций регрессии), следует выбрать семейство, для которого $\hat{\varepsilon}$ минимальна. После этого имеет смысл повторить оценку параметров, используя все доступные данные.

Стоит отметить, что оценки $\hat{\varepsilon}$ обычно консервативны (хуже, чем производительность наилучшего представителя семейства с параметрами, оцененными по всем данным).

Существует несколько подходов к разделению выборки для cross-validation:

- K-fold cross-validation: набор индексов N разбивается на K примерно-равных дизъюнктных подмножества N_k
- Leave-one-out cross-validation: N-fold cross-validation; рассматривается набор N_k, соответствующий всем одноэлементным подмножествам N

3.2 Оценка через комбинаторную размерность

Рассмотрим некоторое параметрическое семейство классификаторов $f_\alpha : \mathbb{R}^p \to \{-1, 1\}$; α — параметры классификатора.

Предполагая, что выборка является выборкой из распределения $\langle x_i, y_i \rangle \sim \mathcal{P}(x, y)$ можно определить риск как

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

Определив эмпирический риск (являющийся случайной величиной) как

$$R_{emp}(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{y_i \neq f_\alpha(x_i)}$$

можно показать, что (независимо от $\mathcal{P}$) с вероятностью $1-\eta$ выполнено неравенство:

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

где h — комбинаторная размерность семейства классификаторов, определяемая как максимальное количество точек, которые при любом их расположении и разделении на классы при некотором α будут безошибочно классифицированны.

Наример, для линейного классификатора в p-мерном пространстве VC-размерность составляет $p+1$, для полиномиального ядра степени d – C_{d+p}^{d} (ввиду числа мономов, являющихся собственными функциями ядра), для RBF VC-размерность бесконечна (см. описанную в 2.1 конструкцию, разделяющую произвольное множество точек на основе RBF).

Можно отметить, что при конечной VC-размерности и $n \to \infty$ дополнительное слагаемое стремится к 0, т.е. $R_{emp}(\alpha) \to R(\alpha)$ по вероятности (и эта сходимость наблюдается вне зависимости от конкретного распределения $\mathcal{P}$).