Линейные методы классификации и регрессии: метод опорных векторов

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: http://shad.yandex.ru/lectures

10 марта 2016

Содержание

Задача обучения линейного классификатора

Дано:

Обучающая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$,

x_i — объекты, векторы из множества $X = \mathbb{R}^n$,

y_i — метки классов, элементы множества $Y = \{-1, +1\}$.

Найти:

Параметры $w \in \mathbb{R}^n$, $w_0 \in \mathbb{R}$ линейной модели классификации

$$a(x; w, w_0) = \operatorname{sign}\bigl(\langle x, w\rangle - w_0\bigr).$$

Критерий — минимизация эмпирического риска:

$$\sum_{i=1}^{\ell} \bigl[a(x_i; w, w_0) \neq y_i\bigr] = \sum_{i=1}^{\ell} \bigl[M_i(w, w_0) < 0\bigr] \to \min_{w, w_0}.$$

где $M_i(w, w_0) = \bigl(\langle x_i, w\rangle - w_0\bigr) y_i$ — *отступ* (margin) объекта x_i,

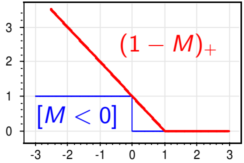
Аппроксимация и регуляризация эмпирического риска

Эмпирический риск — это кусочно-постоянная функция.
Заменим его оценкой сверху, непрерывной по параметрам:

$$Q(w, w_0) = \sum_{i=1}^{\ell} \big[M_i(w, w_0) < 0\big] \leqslant$$

$$\leqslant \sum_{i=1}^{\ell} \big(1 - M_i(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \;\to\; \min_{w, w_0}.$$

- *Аппроксимация* штрафует объекты за приближение к границе классов, увеличивая зазор между классами
- *Регуляризация* штрафует неустойчивые решения в случае мультиколлинеарности



Оптимальная разделяющая гиперплоскость

Линейный классификатор:

$$a(x, w) = \operatorname{sign}\bigl(\langle w, x\rangle - w_0\bigr), \quad w, x \in \mathbb{R}^n,\ w_0 \in \mathbb{R}.$$

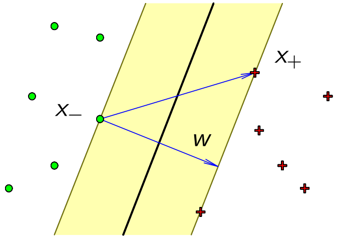
Пусть выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ линейно разделима:

$$\exists w, w_0 : \quad M_i(w, w_0) = y_i\bigl(\langle w, x_i\rangle - w_0\bigr) > 0, \quad i = 1, \ldots, \ell$$

Нормировка: $\min\limits_{i=1,\ldots,\ell} M_i(w, w_0) = 1.$

Разделяющая полоса:

$$\bigl\{x\colon\ -1 \leqslant \langle w, x\rangle - w_0 \leqslant 1\bigr\}.$$



Ширина полосы:

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{2}{\|w\|} \to \max.$$

Обоснование кусочно-линейной функции потерь

Линейно разделимая выборка

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,w_0}; \\ M_i(w,w_0) \geqslant 1, \quad i = 1,\ldots,\ell. \end{cases}$$

Переход к линейно неразделимой выборке (эвристика)

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ M_i(w,w_0) \geqslant 1 - \xi_i, \quad i = 1,\ldots,\ell; \\ \xi_i \geqslant 0, \quad i = 1,\ldots,\ell. \end{cases}$$

Эквивалентная задача безусловной минимизации:

$$C\sum_{i=1}^{\ell}\bigl(1 - M_i(w,w_0)\bigr)_+ + \frac{1}{2}\|w\|^2 \to \min_{w,w_0}.$$

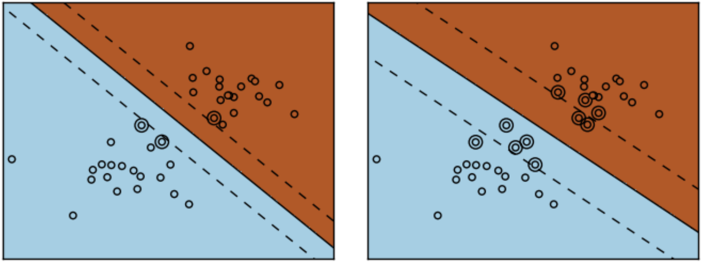
Влияние константы C на решение SVM

SVM — аппроксимация и регуляризация эмпирического риска:

$$\sum_{i=1}^{\ell}\bigl(1 - M_i(w, w_0)\bigr)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w, w_0}.$$

большой C
слабая регуляризация

малый C
сильная регуляризация



Пример из Python scikits learn: `http://scikit-learn.org/dev`

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0;\ h_j(x) = 0;\ \text{(исходные ограничения)} \\ \mu_i \geqslant 0;\ \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0;\ \text{(условие дополняющей нежёсткости)} \end{cases}$$

Применение условий ККТ к задаче SVM

Функция Лагранжа: $\mathscr{L}(w, w_0, \xi; \lambda, \eta) =$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i \bigl(M_i(w, w_0) - 1\bigr) - \sum_{i=1}^{\ell} \xi_i \bigl(\lambda_i + \eta_i - C\bigr),$$

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

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial w} = 0, \quad \dfrac{\partial \mathscr{L}}{\partial w_0} = 0, \quad \dfrac{\partial \mathscr{L}}{\partial \xi} = 0; \\ \xi_i \geqslant 0, \quad \lambda_i \geqslant 0, \quad \eta_i \geqslant 0, \quad i = 1, \ldots, \ell; \\ \lambda_i = 0 \ \text{ либо } \ M_i(w, w_0) = 1 - \xi_i, \quad i = 1, \ldots, \ell; \\ \eta_i = 0 \ \text{ либо } \ \xi_i = 0, \quad i = 1, \ldots, \ell; \end{cases}$$

Необходимые условия седловой точки функции Лагранжа

Функция Лагранжа: $\mathscr{L}(w, w_0, \xi; \lambda, \eta) =$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i \bigl(M_i(w, w_0) - 1\bigr) - \sum_{i=1}^{\ell} \xi_i \bigl(\lambda_i + \eta_i - C\bigr),$$

Необходимые условия седловой точки функции Лагранжа:

$$\frac{\partial \mathscr{L}}{\partial w} = w - \sum_{i=1}^{\ell} \lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell} \lambda_i y_i x_i;$$

$$\frac{\partial \mathscr{L}}{\partial w_0} = -\sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0;$$

$$\frac{\partial \mathscr{L}}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 \quad \Longrightarrow \quad \eta_i + \lambda_i = C, \quad i = 1, \dots, \ell.$$

Понятие опорного вектора

Типизация объектов:

1. $\lambda_i = 0$; $\eta_i = C$; $\xi_i = 0$; $M_i \geqslant 1$.
   — периферийные (неинформативные) объекты.
2. $0 < \lambda_i < C$; $0 < \eta_i < C$; $\xi_i = 0$; $M_i = 1$.
   — опорные граничные объекты.
3. $\lambda_i = C$; $\eta_i = 0$; $\xi_i > 0$; $M_i < 1$.
   — опорные-нарушители.

Определение

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

Двойственная задача

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

Решение прямой задачи выражается через решение двойственной:

$$\begin{cases} w = \sum\limits_{i=1}^{\ell} \lambda_i y_i x_i; \\ w_0 = \langle w, x_i \rangle - y_i, \qquad \text{для любого } i\text{: } \lambda_i > 0,\ M_i = 1. \end{cases}$$

Линейный классификатор:

$$a(x) = \operatorname{sign}\Bigl(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle - w_0 \Bigr).$$

Нелинейное обобщение SVM

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

Определение

Функция $K\colon X \times X \to \mathbb{R}$ — *ядро*, если $K(x, x') = \langle \psi(x), \psi(x') \rangle$
при некотором $\psi\colon X \to H$, 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:
 $\int_X \int_X K(x, x') g(x) g(x') dx dx' \geqslant 0$ для любой $g\colon X \to \mathbb{R}$.

Конструктивные методы синтеза ядер

1. $K(x,x') = \langle x,x' \rangle$ — 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. $\forall\varphi : X{\to}X$ если K_0 ядро, то $K(x,x') = K_0(\varphi(x), \varphi(x'))$ — ядро;
7. если $s\colon X \times X \to \mathbb{R}$ — симметричная интегрируемая функция, то $K(x,x') = \int_X s(x,z)s(x',z)\,dz$ — ядро;
8. если K_0 — ядро и функция $f\colon \mathbb{R} \to \mathbb{R}$ представима в виде сходящегося степенного ряда с неотрицательными коэффициентами, то $K(x,x') = f\bigl(K_0(x,x')\bigr)$ — ядро;

Пример: спрямляющее пространство для квадратичного ядра

Пусть $X = \mathbb{R}^2$, $K(u,v) = \langle u, v\rangle^2$, где $u = (u_1, u_2)$, $v = (v_1, v_2)$.

Задача: найти пространство H и преобразование $\psi\colon X \to H$, при которых $K(x, x') = \langle \psi(x), \psi(x')\rangle_H$.

Разложим квадрат скалярного произведения:

$$
\begin{aligned}
K(u,v) = \langle u, v\rangle^2 &= \langle (u_1, u_2), (v_1, v_2)\rangle^2 = \\
&= (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2 u_1 v_1 u_2 v_2 = \\
&= \langle (u_1^2, u_2^2, \sqrt{2} u_1 u_2), (v_1^2, v_2^2, \sqrt{2} v_1 v_2)\rangle.
\end{aligned}
$$

Таким образом,

$$H = \mathbb{R}^3, \quad \psi\colon (u_1, u_2) \mapsto \bigl(u_1^2, u_2^2, \sqrt{2} u_1 u_2\bigr),$$

Линейной поверхности в пространстве H соответствует квадратичная поверхность в исходном пространстве X.

Примеры ядер

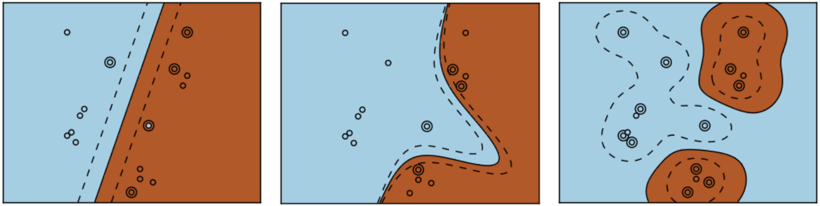
1. $K(x,x') = \langle x,x'\rangle^2$
   — квадратичное ядро;
2. $K(x,x') = \langle x,x'\rangle^d$
   — полиномиальное ядро с мономами степени d;
3. $K(x,x') = \bigl(\langle x,x'\rangle + 1\bigr)^d$
   — полиномиальное ядро с мономами степени $\leqslant d$;
4. $K(x,x') = \sigma\bigl(\langle x,x'\rangle\bigr)$
   — нейросеть с заданной функцией активации $\sigma(z)$
   (не при всех σ является ядром);
5. $K(x,x') = \mathrm{th}\bigl(k_1\langle x,x'\rangle - k_0\bigr)$, $k_0, k_1 \geqslant 0$
   — нейросеть с сигмоидными функциями активации;
6. $K(x,x') = \exp\bigl(-\gamma\|x - x'\|^2\bigr)$
   — сеть радиальных базисных функций (RBF ядро);

Классификация с различными ядрами

Гиперплоскость в спрямляющем пространстве соответствует нелинейной разделяющей поверхности в исходном.

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

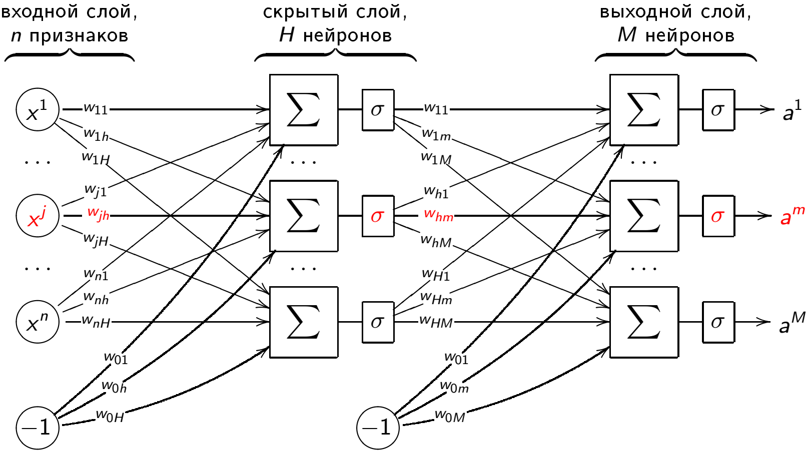
линейное	полиномиальное	гауссовское (RBF)
$\langle x, x' \rangle$	$\left(\langle x, x' \rangle + 1\right)^d, \quad d=3$	$\exp\left(-\gamma \|x - x'\|^2\right)$



Пример из Python scikits learn: `http://scikit-learn.org/dev`

Двухслойная нейронная сеть

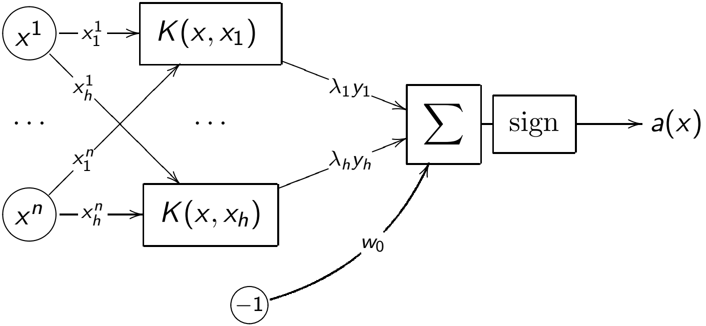
$$a^m(x) = \sigma\left(\sum_{h=0}^{H} w_{hm}\sigma\left(\sum_{j=0}^{J} w_{jh} f_j(x)\right)\right), \quad \sigma \text{ — функция активации}$$



SVM как двухслойная нейронная сеть

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

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{h} \lambda_i y_i K(x, x_i) - w_0\right).$$



Первый слой вместо скалярных произведений вычисляет ядра

Преимущества и недостатки SVM

Преимущества SVM перед SG и нейронными сетями:

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

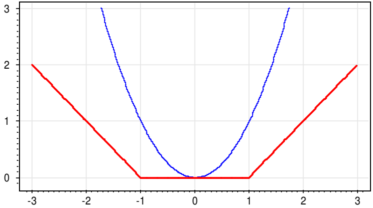
Недостатки классического SVM:

- Нет общих подходов к оптимизации $K(x, x')$ под задачу.
- Нет «встроенного» отбора признаков.
- Приходится подбирать константу C.

SVM-регрессия

Модель регрессии: $a(x) = \langle x, w \rangle - w_0$, $\;w \in \mathbb{R}^n$, $\;w_0 \in \mathbb{R}$.

Функция потерь: $\mathscr{L}(\varepsilon) = \bigl(|\varepsilon| - \delta\bigr)_+$ в сравнении с $\mathscr{L}(\varepsilon) = \varepsilon^2$:



Постановка задачи:

$$\sum_{i=1}^{\ell} \bigl(|\langle w, x_i \rangle - w_0 - y_i| - \delta\bigr)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w, w_0}.$$

Задача решается путём замены переменных
и сведения к задаче квадратичного программирования

SVM-регрессия

Замена переменных:

$$\xi_i^+ = \bigl(\langle w, x_i\rangle - w_0 - y_i - \delta\bigr)_+;$$
$$\xi_i^- = \bigl(-\langle w, x_i\rangle + w_0 + y_i - \delta\bigr)_+;$$

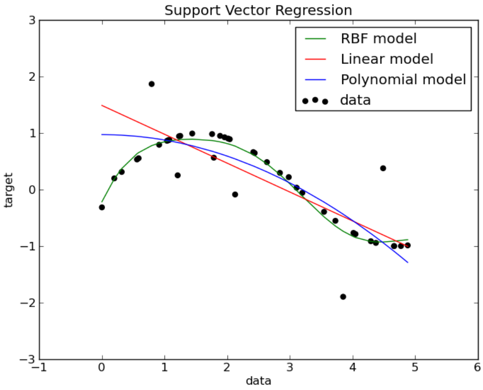
Постановка задачи SVM-регрессии:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}(\xi_i^+ + \xi_i^-) \to \min\limits_{w, w_0, \xi^+, \xi^-}; \\ y_i - \delta - \xi_i^- \leqslant \langle w, x_i\rangle - w_0 \leqslant y_i + \delta + \xi_i^+, \quad i = 1, \ldots, \ell; \\ \xi_i^- \geqslant 0, \quad \xi_i^+ \geqslant 0, \quad i = 1, \ldots, \ell. \end{cases}$$

Это задача квадратичного программирования
с линейными ограничениями-неравенствами,
решается также сведением к двойственной задаче.

Пример из Python scikits learn

Сравнение SVM-регрессии с гауссовским (RBF) ядром, линейной и полиномиальной регрессией:



---

http://scikit-learn.org/0.5/auto_examples/svm/plot_svm_regression.html

1-norm SVM (LASSO SVM)

Аппроксимация эмпирического риска с L_1-регуляризацией:

$$\sum_{i=1}^{\ell}\bigl(1-M_i(w,w_0)\bigr)_+ + \mu\sum_{j=1}^{n}|w_j| \;\to\; \min_{w,w_0}.$$

- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊖ LASSO начинает отбрасывать значимые признаки, когда ещё не все шумовые отброшены
- ⊖ Нет *эффекта группировки* (grouping effect): значимые зависимые признаки должны отбираться вместе и иметь примерно равные веса w_j

---

Bradley P., Mangasarian O. Feature selection via concave minimization and support vector machines // ICML 1998.

1-norm SVM (LASSO SVM)

Аппроксимация эмпирического риска с L_1-регуляризацией:

$$\sum_{i=1}^{\ell}\bigl(1 - M_i(w, w_0)\bigr)_+ + \mu\sum_{j=1}^{n}|w_j| \;\to\; \min_{w,w_0}.$$

Почему L_1-регуляризатор приводит к отбору признаков?

Замена переменных: $u_j = \frac{1}{2}\bigl(|w_j| + w_j\bigr)$, $v_j = \frac{1}{2}\bigl(|w_j| - w_j\bigr)$.
Тогда $w_j = u_j - v_j$ и $|w_j| = u_j + v_j$;

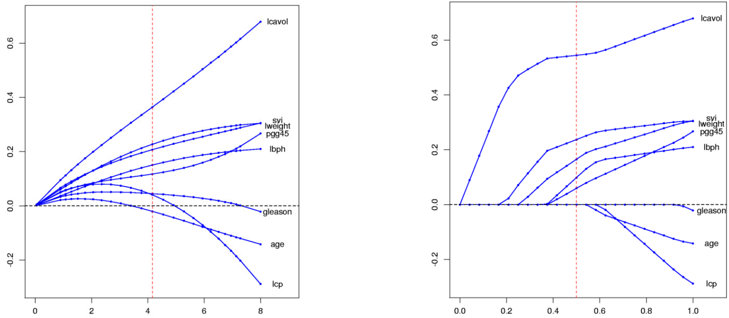
$$\begin{cases}\displaystyle\sum_{i=1}^{\ell}\bigl(1 - M_i(u - v, w_0)\bigr)_+ + \mu\sum_{j=1}^{n}(u_j + v_j) \to \min_{u,v}\\ u_j \geqslant 0, \quad v_j \geqslant 0, \quad j = 1, \dots, n;\end{cases}$$

чем больше μ, тем больше индексов j таких, что $u_j = v_j = 0$, но тогда $w_j = 0$, значит, признак не учитывается.

Сравнение L_2 и L_1 регуляризации

Зависимость весов w_j от коэффициента $\frac{1}{\mu}$

L_2 регуляризатор: $\mu \sum_j w_j^2$ L_1 регуляризатор: $\mu \sum_j |w_j|$



Задача из UCI: prostate cancer (диагностика рака)

T.Hastie, R.Tibshirani, J.Friedman. The Elements of Statistical Learning. Springer, 2001.

Doubly Regularized SVM (Elastic Net SVM)

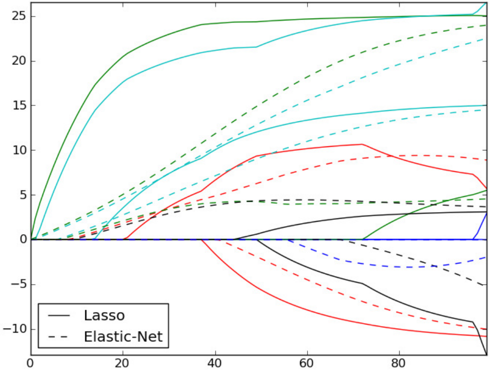
$$C\sum_{i=1}^{\ell}\bigl(1 - M_i(w, w_0)\bigr)_+ + \mu\sum_{j=1}^{n}|w_j| + \frac{1}{2}\sum_{j=1}^{n} w_j^2 \;\to\; \min_{w,w_0}.$$

- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊕ Есть эффект группировки
- ⊖ Шумовые признаки также группируются вместе, и группы значимых признаков могут отбрасываться, когда ещё не все шумовые отброшены

Li Wang, Ji Zhu, Hui Zou. The doubly regularized support vector machine // Statistica Sinica, 2006. No. 16, Pp. 589–615.

Doubly Regularized SVM (Elastic Net SVM)

Elastic Net менее жёстко отбирает признаки.
Зависимости весов w_j от коэффициента $\log \frac{1}{\mu}$:

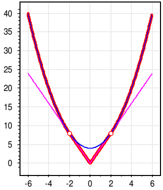


Пример из Python scikits learn:
`scikit-learn.org/0.5/auto_examples/glm/plot_lasso_coordinate_descent_path.html`

Support Features Machine (SFM)

$$C\sum_{i=1}^{\ell}\bigl(1-M_i(w,w_0)\bigr)_+ + \sum_{j=1}^{n} R_\mu(w_j) \;\to\; \min_{w,w_0}.$$

$$R_\mu(w_j) = \begin{cases} 2\mu|w_j|, & |w_j| \leqslant \mu; \\ \mu^2 + w_j^2, & |w_j| \geqslant \mu; \end{cases}$$



- Отбор признаков с параметром *селективности* μ
- Есть эффект группировки
- Значимые зависимые признаки ($|w_j| > \mu$) группируются и входят в решение совместно (как в Elastic Net),
- Шумовые признаки ($|w_j| < \mu$) подавляются независимо (как в LASSO)

Tatarchuk A., Urlov E., Mottl V., Windridge D. A support kernel machine for supervised selective combining of diverse pattern-recognition modalities // Multiple Classifier Systems. LNCS, Springer-Verlag, 2010. Pp. 165–174.

Relevance Features Machine (RFM)

$$C \sum_{i=1}^{\ell} \bigl(1 - M_i(w, w_0)\bigr)_+ + \sum_{j=1}^{n} \ln\Bigl(w_j^2 + \tfrac{1}{\mu}\Bigr) \;\to\; \min_{w, w_0} .$$

- Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- Есть эффект группировки
- Лучше отбирает набор значимых признаков, когда они только совместно обеспечивают хорошее решение

Tatarchuk A., Mottl V., Eliseyev A., Windridge D. Selectivity supervision in combining pattern recognition modalities by feature- and kernel-selective Support Vector Machines // 19th International Conference on Pattern Recognition, Vol 1-6, 2008, Pp. 2336–2339.

Метод релевантных векторов RVM (Relevance Vector Machine)

Положим, как и в SVM, при некоторых $\lambda_i \geqslant 0$

$$w = \sum_{i=1}^{\ell} \lambda_i y_i x_i,$$

причём опорным векторам x_i соответствуют $\lambda_i \neq 0$.

Проблема: Какие из коэффициентов λ_i лучше обнулить?

Идея: пусть регуляризатор зависит не от w, а от λ_i.
Пусть λ_i независимые, гауссовские, с дисперсиями α_i:

$$p(\lambda) = \frac{1}{(2\pi)^{\ell/2}\sqrt{\alpha_1 \cdots \alpha_\ell}} \exp\left(-\sum_{i=1}^{\ell} \frac{\lambda_i^2}{2\alpha_i}\right);$$

$$\sum_{i=1}^{\ell} \bigl(1 - M_i(w(\lambda), w_0)\bigr)_+ + \frac{1}{2}\sum_{i=1}^{\ell}\left(\ln \alpha_i + \frac{\lambda_i^2}{\alpha_i}\right) \to \min_{\lambda,\alpha}.$$

Преимущества и недостатки RVM

Преимущества:

- ⊕ Опорных векторов, как правило, меньше (более «разреженное» решение).
- ⊕ Шумовые выбросы уже не входят в число опорных.
- ⊕ Не надо искать параметр регуляризации (вместо этого α_i оптимизируются в процессе обучения).
- ⊕ Аналогично SVM, можно использовать ядра.

Недостатки:

- ⊖ Авторам не удалось показать практическое преимущество по качеству классификации.

Резюме по линейным классификаторам

- *SVM* — лучший метод линейной классификации
- *SVM* изящно обобщается для нелинейной классификации, для линейной и нелинейной регрессии
- *Аппроксимация пороговой функции потерь* $\mathscr{L}(M)$ увеличивает зазор и повышает качество классификации
- *Регуляризация* устраняет мультиколлинеарность и переобучение
- *Регуляризация* эквивалентна введению априорного распределения в пространстве коэффициентов
- L_1 и другие нестандартные регуляризаторы делают отбор признаков без явного перебора подмножеств