Линейные методы классификации: метод опорных векторов

К. В. Воронцов
vokov@forecsys.ru

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

март 2014

Содержание

Задача SVM — Support Vector Machine

Задача классификации: $X = \mathbb{R}^n$, $Y = \{-1, +1\}$,
по обучающей выборке $X^\ell = (x_i, y_i)_{i=1}^\ell$
найти параметры $w \in \mathbb{R}^n$, $w_0 \in \mathbb{R}$ алгоритма классификации

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

Метод минимизации аппроксимированного регуляризованного эмпирического риска:

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

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

Почему именно такая функция потерь? и такой регуляризатор?

Оптимальная разделяющая гиперплоскость

Линейный классификатор:

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

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

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

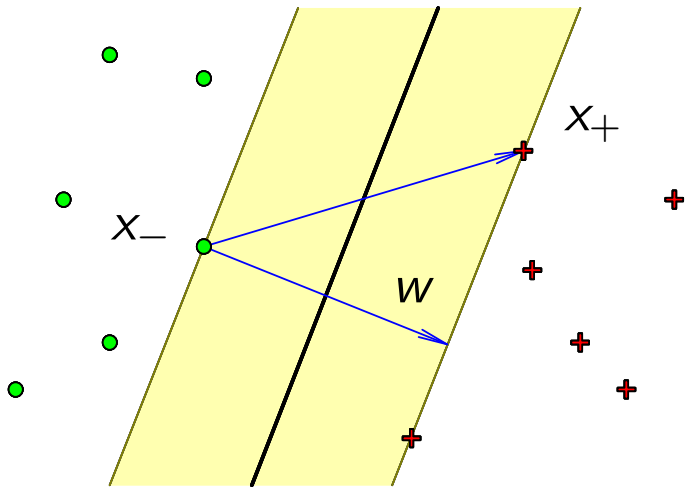
Нормировка: $\min\limits_{i=1,\ldots,\ell} M_i(w, w_0) = 1.$

Разделяющая полоса:

$$\big\{x:\ -1 \leqslant \langle w, x\rangle - w_0 \leqslant 1\big\}.$$

Ширина полосы:

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



Обоснование кусочно-линейной функции потерь

Линейно разделимая выборка

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

Переход к линейно неразделимой выборке (эвристика)

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

Эквивалентная задача безусловной минимизации:

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

Напоминание. Условия Каруша–Куна–Таккера

Задача математического программирования:

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

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

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

Применение условий KKT к задаче SVM

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

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

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

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

Необходимые условия седловой точки функции Лагранжа:

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

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

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

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

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

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:

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

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: $\int_X \int_X K(x,x')g(x)g(x')dxdx' \geqslant 0$ 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 $f : \mathbb{R} \to \mathbb{R}$ представима в виде сходящегося степенного ряда с неотрицательными коэффициентами, то $K(x, x') = f\big(K_0(x, x')\big)$ — 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:

$$\begin{aligned} K(u,v) = \langle u,v\rangle^2 &= \langle (u_1,u_2),(v_1,v_2)\rangle^2 = \\ &= (u_1v_1 + u_2v_2)^2 = u_1^2v_1^2 + u_2^2v_2^2 + 2u_1v_1u_2v_2 = \\ &= \langle (u_1^2, u_2^2, \sqrt{2}u_1u_2), (v_1^2, v_2^2, \sqrt{2}v_1v_2)\rangle. \end{aligned}$$

Таким образом,

$$H = \mathbb{R}^3, \quad \psi\colon (u_1,u_2) \mapsto \left(u_1^2, u_2^2, \sqrt{2}u_1u_2\right),$$

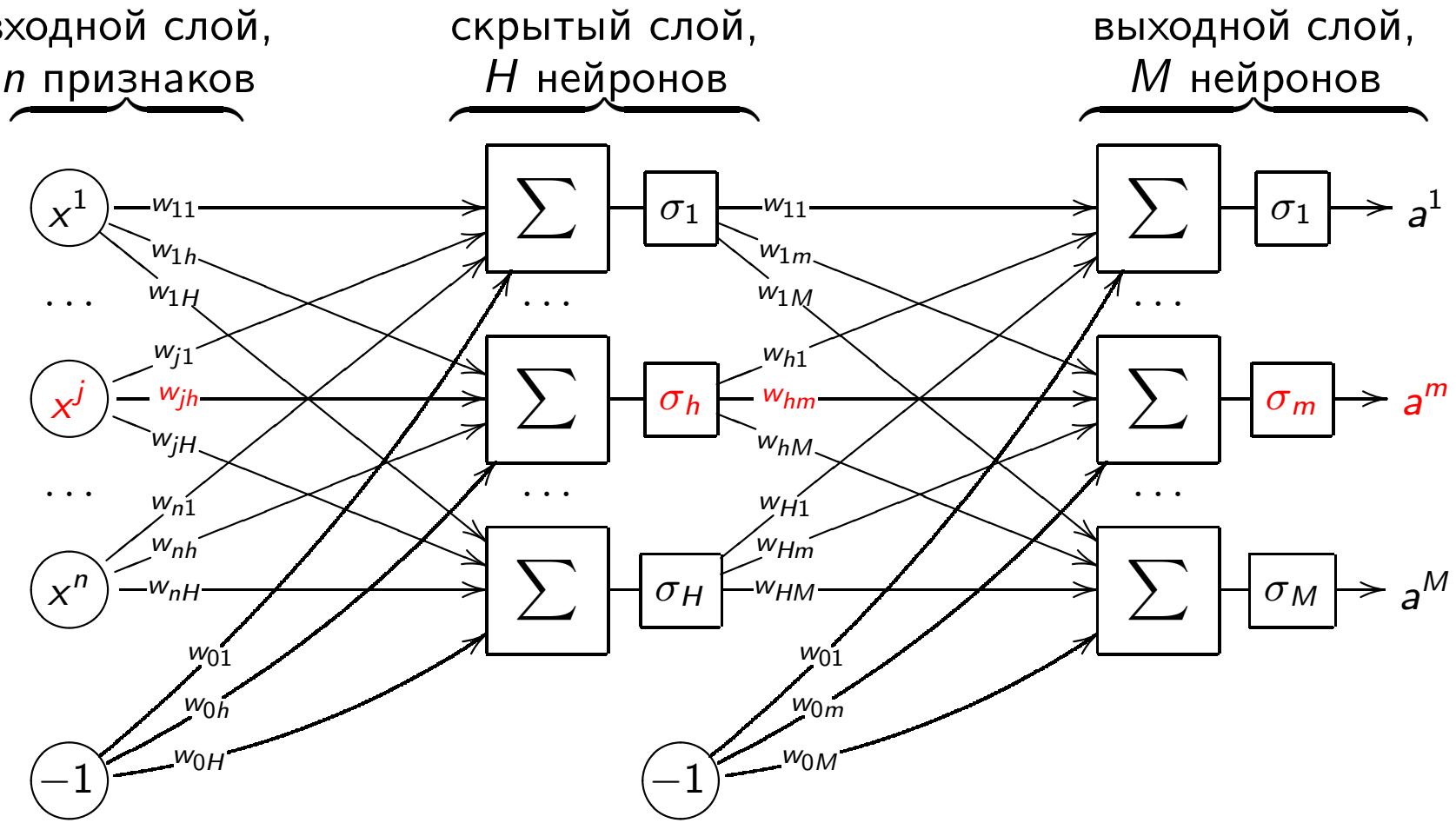
Линейной поверхности в пространстве H соответствует квадратичная поверхность в исходном пространстве X.

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

Двухслойная нейронная сеть

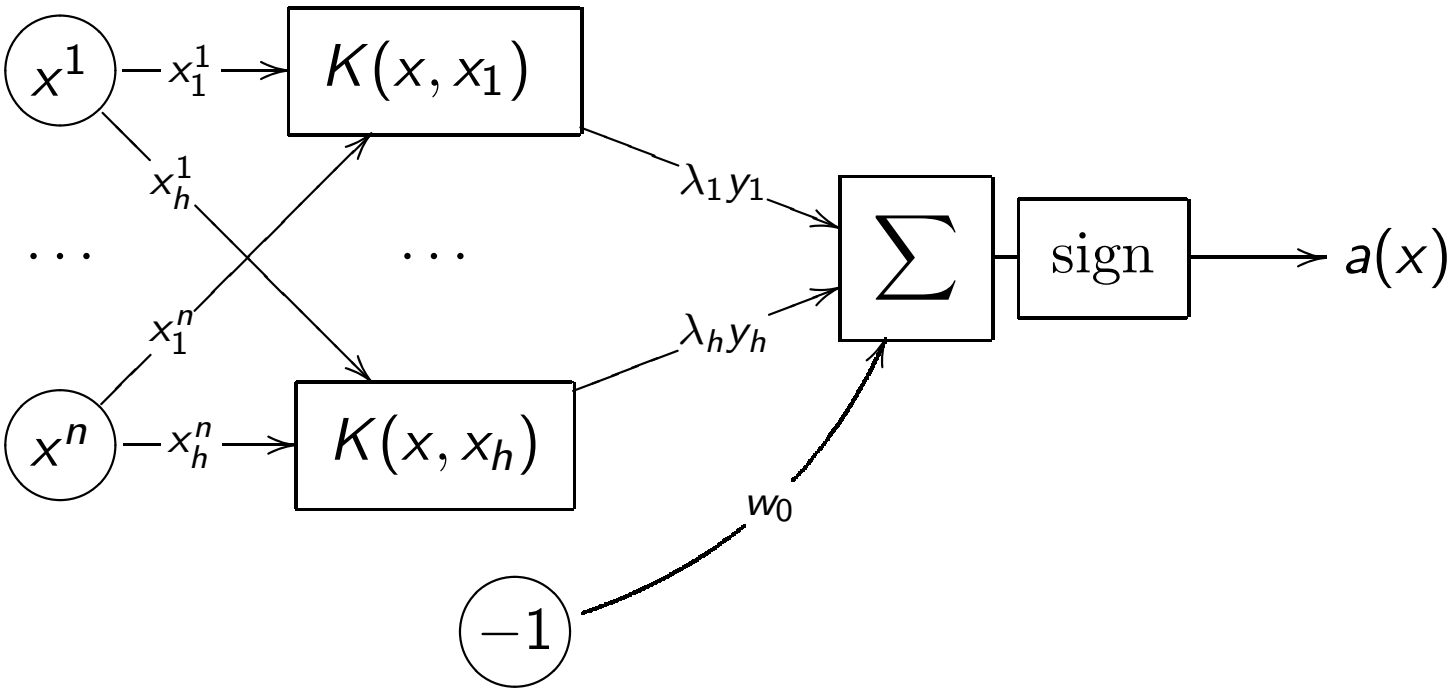
$$a^m(x) = \sigma_m\left(\sum_{h=0}^{H} w_{hm}\sigma_h\left(\sum_{j=0}^{J} w_{jh} f_j(x)\right)\right).$$



SVM как двухслойная нейронная сеть

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

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



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:

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

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:

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

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

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.

1-norm SVM (LASSO SVM)

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

- ⊕ 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.

Doubly Regularized SVM (Elastic Net SVM)

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

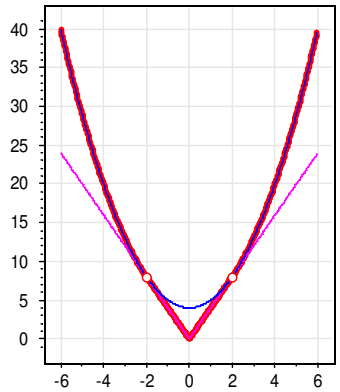
- ⊕ Отбор признаков с параметром *селективности* μ: чем больше μ, тем меньше признаков останется
- ⊕ Есть эффект группировки
- ⊖ Шумовые признаки также группируются вместе, и группы значимых признаков могут отбрасываться, когда ещё не все шумовые отброшены

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

Support Features Machine (SFM)

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

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



- 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)

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

- 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.

Балансировка ошибок I и II рода

Задача классификации на два класса, $Y = \{-1, +1\}$;
Модель классификации: $a(x, w, w_0) = \text{sign}\big(f(x, w) - w_0\big)$.

$a(x_i, w) = -1$, $y_i = +1$ — ложно-отрицательная классификация («пропуск цели», ошибка I рода)

$a(x_i, w) = +1$, $y_i = -1$ — ложно-положительная классификация («ложная тревога», ошибка II рода)

На практике цена ошибок I и II рода может быть неизвестна или многократно пересматриваться.

Постановка задачи

- Выбирать w_0 без обучения w заново.
- Ввести характеристику качества классификатора, инвариантную относительно выбора цены ошибок.

Определение ROC-кривой

ROC — «receiver operating characteristic».

- Каждая точка кривой соответствует некоторому $a(x; w, w_0)$.
- по оси X: доля *ложно-положительных классификаций* (FPR — false positive rate):

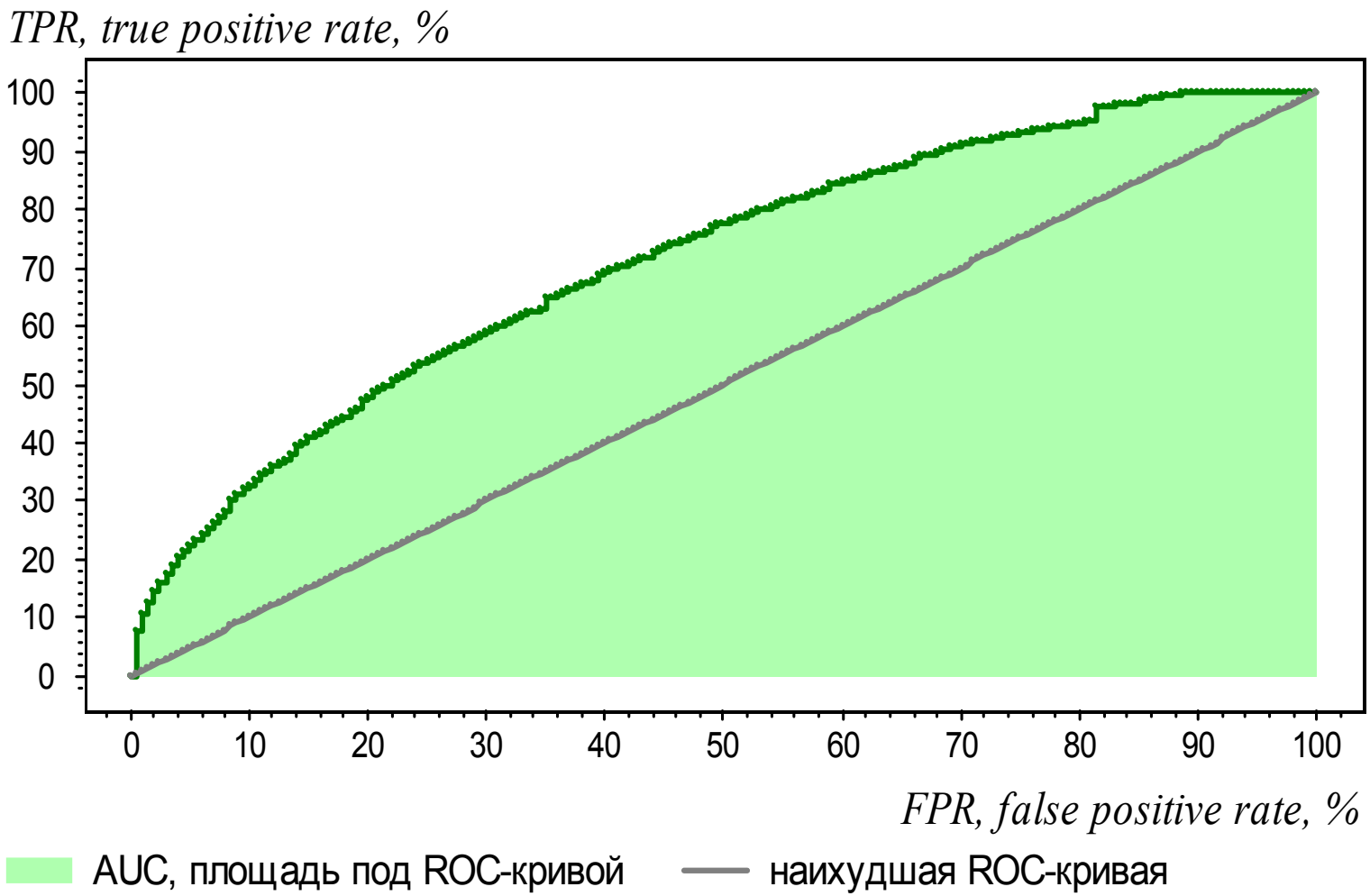
$$\text{FPR}(a, X^\ell) = \frac{\sum_{i=1}^{\ell} [y_i = -1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell} [y_i = -1]};$$

  $1 - \text{FPR}(a)$ называется *специфичностью* алгоритма a.
- по оси Y: доля *верно-положительных классификаций* (TPR — true positive rate):

$$\text{TPR}(a, X^\ell) = \frac{\sum_{i=1}^{\ell} [y_i = +1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell} [y_i = +1]};$$

  $\text{TPR}(a)$ называется также *чувствительностью* алгоритма a.

Пример ROC-кривой



Алгоритм эффективного построения ROC-кривой

Вход: выборка X^ℓ; дискриминантная функция $f(x, w)$;

Выход: $\{(\mathrm{FPR}_i, \mathrm{TPR}_i)\}_{i=0}^{\ell}$, AUC — площадь под ROC-кривой.

$\ell_y := \sum_{i=1}^{\ell} [y_i = y]$, для всех $y \in Y$;

упорядочить выборку X^ℓ по убыванию значений $f(x_i, w)$;

поставить первую точку в начало координат:
$(\mathrm{FPR}_0, \mathrm{TPR}_0) := (0, 0)$; $\mathrm{AUC} := 0$;

для $i := 1, \ldots, \ell$

  если $y_i = -1$ **то** сместиться на один шаг вправо:

    $\mathrm{FPR}_i := \mathrm{FPR}_{i-1} + \frac{1}{\ell_-}$; $\mathrm{TPR}_i := \mathrm{TPR}_{i-1}$;

    $\mathrm{AUC} := \mathrm{AUC} + \frac{1}{\ell_-}\mathrm{TPR}_i$;

  иначе сместиться на один шаг вверх:

    $\mathrm{FPR}_i := \mathrm{FPR}_{i-1}$; $\mathrm{TPR}_i := \mathrm{TPR}_{i-1} + \frac{1}{\ell_+}$;

Градиентная максимизация AUC

Модель: $a(x_i, w, w_0) = \text{sign}\big(f(x_i, w) - w_0\big)$.

AUC — это доля правильно упорядоченных пар (x_i, x_j):

$$\text{AUC} = \frac{1}{\ell_-}\sum_{i=1}^{\ell}\big[y_i = -1\big]\,\text{TPR}_i = $$

$$= \frac{1}{\ell_-\ell_+}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\big[y_i < y_j\big]\big[f(x_i, w) < f(x_j, w)\big] \to \max_w.$$

Явная максимизация аппроксимированного AUC:

$$Q(w) = \sum_{i,j:\ y_i<y_j} \mathscr{L}\big(\underbrace{f(x_j, w) - f(x_i, w)}_{M_{ij}(w)}\big) \to \min_w,$$

где $\mathscr{L}(M)$ — убывающая функция отступа,
$M_{ij}(w)$ — новое понятие отступа для пар объектов.

Резюме по линейным классификаторам

- Методы обучения линейных классификаторов отличаются
  — видом функции потерь;
  — видом регуляризатора;
  — численным методом оптимизации.
- *Аппроксимация пороговой функции потерь* гладкой убывающей функцией отступа $\mathscr{L}(M)$ повышает качество классификации (за счёт увеличения зазора)
  и облегчает оптимизацию.
- *Регуляризация* решает проблему мультиколлинеарности и также снижает переобучение.
- *SVM* — один из лучших методов машинного обучения, изящно обобщается для нелинейной классификации
- *AUC* инвариантна относительно выбора цены ошибок.