Classification: Support Vector Classifier

Xiang Zhou

School of Data Science
Department of Mathematics
City University of Hong Kong

SVM

developed by computer science

- Vapnik 1995: Geometric Viewpoint + Primal-Dual for Quadratic Programming (+ Kernel trick, new def of metric)
- Sollich 2002: Bayesian Viewpoint

Method	main properties
maximal margin classifier	only for linear separable dataset
support vector classifier	slack variable, linear classifier
support vector machine	kernel trick, nonlinear classifier

Table: Development of SVM

1

[1]We do not discuss here the numerical optimization part of SVM (a good example for convex optimization . online resource: http:/uito/www.robots.ox.ac.uk/~az/lectures/ml/index.html). The focus here is the geometric intuition and modelling.

Linear Separable Problem

Binary classification problem: dataset $\{x_i, y_i\}$ where $y_i \in \mathcal{Y} = \{-1, 1\}$. Recall

- Logistic regression assumes: log odd $\log h(x)$ is linear in x. The decision boudary $h(x) = 0.5$ is equivalent to $\beta \cdot x = 0.5$
- The LDA's the discriminant function $\delta(x)$ is also linear in x.
- SVM is also a linear classifier, with a strong geometric intuition.

Remark

- *The logistic regression = sigmoid activation function + linear feature assumption + maximum likelihood*
- *The linear discriminant analysis (LDA) = Bayes classifier + Gaussian mixture + equal variance assumption*
- *The support vector machine (SVM) = linear classifier + max margin*

Note the notations different from logistic regressions:

- $\mathcal{Y} = \{-1, 1\}$, not $\{0, 1\}$
- the discriminant function is generally denoted by f. The classifier $\phi(x) := \text{sign} f(x) \in \{-1, 1\}$. Then decision boundary is $f(x) = 0$, not $h(x) = 0.5$.

This set of notation is convenient because if y belong to $\{-1, 1\}$

$$\boxed{\text{sign} f(x) = y \iff y f(x) > 0.}$$

Remember $\text{sign} f(x) = \text{sign}(\lambda f(x))$ for any $\lambda > 0$.

The 0-1 loss then can be written as

$$\ell_{01}(f(x), y) = 1 - \text{heaviside}(y f(x)) = (1 - \text{sign}(y f(x)))/2$$

which is equal to $\ell_{01}(\phi(x), y) = \ell_{01}(\text{sign} f(x), y)$. We extend ℓ_{01}'s domain $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R} \times \mathbb{R}$.

Exercise

A linear discriminant function is $f(x) = w \cdot x + b$. *Only the sign matters, so w.l.o.g., we assume* $\|w\| = 1$. *Given a point* x^*, *show the signed distance between* x^* *and the hyperplane* $f(x) = 0$ *is*

$$f(x^*)$$

(or $f(x^*)/\|w\|$ *in general).*

Given one data example (x_i, y_i), if f correctly classifies x_i, then $\operatorname{sign} f(x_i) = y_i$ the distance to the hyperplane $f(x) = 0$ is

$$|f(x_i)| = f(x_i) \cdot \operatorname{sign} f(x_i) = \boxed{f(x_i) y_i} =: M_i,$$

which is the **margin** from x_i to the separating hyperplane.

definition (margin)

Given the dataset $(x_i, y_i), i = 1, \ldots, n$ *and a linear function* $f(x) = w \cdot x + b$*, then the margin of the dataset* $(x_i, y_i), i = 1, \ldots, n$ *to the hyperplane* $f(x) = 0$ *is*

$$M = \min_{1 \le i \le n} \{ y_i (w \cdot x_i + b) / \|w\| \}$$

The support vectors are the collection of $\{x_j\}$ *such that* $M = y_j (w \cdot x_j + b)$*. Sometimes, the margin refers to the two hyperplanes* $w \cdot x + b = \pm M / \|w\|$ *where support vectors lie.*

$M > 0 \iff$ the dataset is linearly separable, i.e.

$$\operatorname{sign} f(x_i) = y_i, \forall i.$$

Definition (maximal margin classifier)

The maximal margin classifier solves the problem

$$\begin{aligned} \max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad & M \\ \text{subject to} \quad & \|w\| = 1 \\ & y_i(w\cdot x_i + b) \geq M, \forall i \end{aligned} \tag{1}$$

- The equivalent form of maximal margin classifier is

$$\begin{aligned} \max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad & M \\ \text{subject to} \quad & y_i(w\cdot x_i + b)/\|w\| \geq M, \quad \forall i \end{aligned} \tag{2}$$

- The constraint $\|w\| = 1$ is only for the uniqueness of w and b; without this constraint, the solution is a family of the linear discriminant functions $\{\lambda f^*(x) : \lambda > 0\}$, which all share the **same** classifier $\phi^* = \text{sign} f$.
- This form is applicable to non linear separable case. If the maximal M is negative, then the dataset is not linearly separable. Otherwise, the dataset is linearly separable.

Exercise (XOR)

Suppose the dataset has $n = 4$ examples as follows:

$$\begin{array}{ll} x_1=(1,\ -1) & y_1 = -1 \\ x_2=(1,\ 1) & y_2 = 1 \\ x_3=(-1,\ 1) & y_3 = -1 \\ x_4=(-1,\ -1) & y_4 = 1 \end{array}$$

. Find the maximal margin classifier

$f(x) = w_1 x_{(1)} + w_2 x_{(2)} + b.$

$$\max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad M$$

$$\begin{aligned} \text{subject to} \quad & w_1^2 + w_2^2 = 1 \\ & w_1 + w_2 + b \geq M \\ & -w_1 - w_2 + b \geq M \\ & w_1 - w_2 - b \geq M \\ & -w_1 + w_2 - b \geq M \end{aligned}$$

The constraints are equivalent to $|w_1 + w_2| \leq -M + b$ and $|w_1 - w_2| \leq -M - b$. Then $|w_1| \leq -M$. So any admissible M is negative. It is easy to show that $M \pm b \leq 0$. So the possible max of M is $M = b$ or $M = -b$. If $M = b$, then $w_1 = -w_2 = \pm b$ and $f(x) = (-x_1 + x_2 + 1)/\sqrt{2}$. If $M = -b$, then $w_1 = w_2 = \pm b$ and the solution is

The alternative form of maximal margin classifier is

$$\max_{w\in\mathbb{R}^d, b\in\mathbb{R}} M$$

$$\text{subject to} \quad y_i(w\cdot x_i + b)/\left\|w\right\| \geq M, \quad \forall i$$

Since we can scale w, b by a **positive** factor arbitrarily, we can assume $M > 0$ and $M\left\|w\right\| = 1$ *if the dataset is linearly separable*, instead of using the rescaling $\left\|w\right\| = 1$. Then

$$\begin{aligned} \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \ & \frac{1}{2}\left\|w\right\|^2 \\ \text{subject to} \ & \boxed{y_i(w\cdot x_i + b) \geq 1, \forall i} \end{aligned} \tag{3}$$

- Now there is NO solution if not linear separable, in contrast to (2) and (1).
- The problem (3) is the standard quadratic programming problem ☺, in contrast to (2) and (1).
- The margin corresponds to the equalities when the inequality constraint, i.e., the two parallel hyperplanes for the margin are given $\boxed{w\cdot x + b = \pm 1}$.

  The margin width is $\frac{2}{\|w\|}$

Support Vector Classifier
soft margin and slack variable

But linear separation assumption is too strong in practice

The non-separable case means there are some examples (x_m, y_m) such that $y_m(w \cdot x_m + b) < 0$. Then by adding n slack variables $\xi = (\xi_1, \ldots, \xi_n)$, we have the support vector classifier

Definition (support vector classifier)

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|^2 \tag{4}$$

$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i \tag{5}$$

$$\xi_i \geq 0, \forall i \tag{6}$$

$$\sum_{i=1}^{n} \xi_i \leq const \tag{7}$$

where $const > 0$ is a tuning parameter.

$const = 0 \iff$ maximal margin classifier (for linear separable case), do

Understand SVC's geometric perspective

- The margin is given by two hyperplanes : $w \cdot x + b = \pm 1$ with the margin gap $2M = \frac{2}{\|w\|}$.
- $\xi_i > 1$ means $y_i(w \cdot x_i + b)$ is negative: y_i is on the other side of the hyperplane predicted by $f(x)$.
- $\xi_i > 0$ then y_i violates the margin;
- $\xi_i = 0$, then y_i is on the same side predicted by the margin; Furthermore, $y_i(w \cdot x_i + b) = 1 \iff$ support vectors

Note that $y_i f(x_i) \geq 1 - \xi_i$ and $\xi_i \geq 0$ together are equivalent to $\xi_i \geq \max\{0, 1 - y_i f(x_i)\} =: (1 - y_i f(x_i))_+$. Then the SVC

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|^2$$

$$\text{subject to} \;\; \xi_i \geq (1 - y_i(w \cdot x_i + b))_+, \forall i$$

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

is equivalent to

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2C} \|w\|^2 + \sum_i \xi_i$$

$$\text{subject to} \;\; \xi_i \geq (1 - y_i(w \cdot x_i + b))_+, \forall i$$

which is equivalent to

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2C} \|w\|^2 + \sum_i (1 - y_i(w \cdot x_i + b))_+ \tag{8}$$

This is the form of (hinge) loss + (L_2) regularization

SVC : hinge Loss + Regularization

$$\min_{w,b} \sum_{i=1}^{n} \ell_{\textsf{hinge}}(y_i, f(x_i)) + \frac{C}{2} \|w\|^2 \tag{9}$$

where $f(x) = w \cdot x + b$ and

$$\ell_{\textsf{hinge}}(y, f) = (1 - yf)_+, \quad y \in \{-1, 1\}, f \in \mathbb{R}$$

$C = \infty \iff w = 0$;
$C = 0 \implies$ (1) min=0 means the linear separation case (2) min > 0 is the non separable; in both, the solution f^* is not unique, even restricted to linear.

logistic regression : binomial deviance Loss without Regularization

Recall the logistic regression solves

$$\min_f \mathbb{E}\, \ell_{\mathsf{bd}}(Y, f(X)) \approx \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathsf{bd}}(y_i, f(x_i)) \tag{10}$$

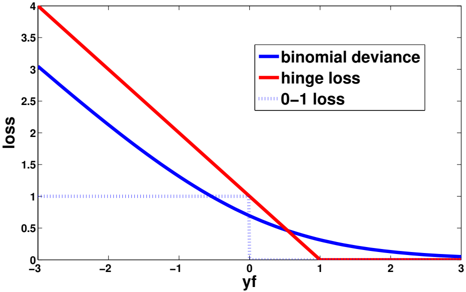
where $f(x) = \text{logit}(h) = \log \frac{h}{1-h}$ with $h(x) = \mathbb{P}(Y = +1|X = x)$ and the binomial deviance loss

$$\ell_{\mathsf{bd}}(y, f) = \log(1 + e^{-yf}), \quad y \in \{-1, 1\}\,.$$

Recall the 0-1 loss (??) in the Bayesian classifier, we rewrite it in term of f:

$$\ell_{01}(y,f) = \mathbf{1}(y \neq \text{sign}(f(x))) = \begin{cases} 1 & \text{if } yf(x) < 0 \\ 0 & \text{if } yf(x) > 0 \end{cases} = 1 - \text{Heaviside}(yf).$$

Then we have three loss functions ℓ_{bd}, ℓ_{hinge}, ℓ_{01} which are all functions effectively in term of the product $yf(x)$



discussion: What differences ? Computational issues ? Which data examples feel the "gradient" force? Why need regularization for hinge? What else of loss function do you like to propose ?

We already know that the optimal solution to the 0-1 loss

$$\inf_f \mathbb{E}\, \ell_{0,1}(Y, f(X))$$

is Bayesian classifier $\phi^*(x) = \text{sign}(f^*) = \text{sign}(h(x) - 0.5)$ where $h(x) = \mathbb{P}(Y = +1|X = x)$. Only the sign f^* is determined.

Exercise

Consider the minimization problem

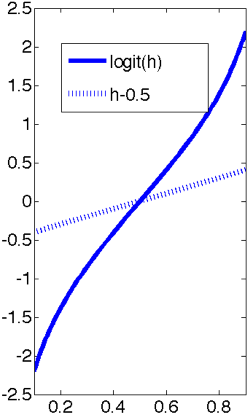
$$\inf_f \mathbb{E}\, \ell_{\textit{bd}}(Y, f(X))$$

for the $\{\pm 1\}$-encoded binary classification problem. Show that the optimal f^* is the log odd:

$$f^*(x) = \textit{logit}(h(x)) = \log \frac{h}{1-h}$$

where $h(x) = \mathbb{P}(Y = +1|X = x)$.

The two problems are not variation of calculus, but are solved in point-wise sense.


logit(h)
h-0.5
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
0.2
0.4
0.6
0.8

Kernel logistic regression vs Kernel SVM

https://stats.stackexchange.com/questions/43996/kernel-logistic-regression-vs-svm