Classification: Support Vector Machine

Xiang Zhou

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

SVM: support vector machine

Mainly developed and was dominantly hot in computer science / pattern recognition

- Vapnik 1995: Geometric Viewpoint + Primal-Dual for Quadratic Programming (+ Kernel trick, new def of metric)
- Major developments throughout 1990's
- Has good generalization properties;
- One of the most important and successful developments before deep learning.

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

2

[1] Sollich 2002: Bayesian Viewpoint of SVM

Linear Separation for binary classification

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

- 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.

Summary:

- 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 $\{-1, 1\}$

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

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

Exercise

A linear discriminant function is $f(x) = w \cdot x + b$. 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). The positive sign of $f(x^*)$ means that x^* on the same side of the hyperplane as the normal direction vector w.

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

$$\frac{1}{\|w\|}|f(x_i)| = \frac{1}{\|w\|} f(x_i) \cdot \mathrm{sign} f(x_i) = \boxed{\frac{1}{\|w\|} f(x_i) y_i =: M_i},$$

which is the **margin** of x_i and y_i to the separating hyperplane. If f misclassifies x_i, then $M_i < 0$.

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 [a] of the dataset $(x_i, y_i), i = 1, \ldots, n$ to the hyperplane $f(x) = 0$ is

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

The **support vectors** are the collection of $\{x_j\}$ such that $M = y_j (w \cdot x_j + b)$.

[a] Sometimes, the margin refers to $2M$, the distance between the two hyperplanes $w \cdot x + b = \pm M / \|w\|$.

- M depends on w, b and the dataset $\{x_i, y_i\}$.
- If there exists a linear function $f = w \cdot x + b$ such that $M > 0 \iff$ the dataset is linearly separable, i.e.

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

- The margin M is a function of w and b, we consider its maximal value over the choice of w and b:

$$M^* := \max_{w,b} M(w,b) = \max_{w,b} \left(\min_i \{y_i(w \cdot x_i + b) / \|w\|\} \right) \tag{1}$$

- If M^* is positive, then the corresponding optimal w^* and b^* : the dataset is linearly separabble
- If M^* is negative, then for any w and b, $M < 0$, i.e., there exist some data i such that $M_i < 0$, misclassified. The dataset is not linearly separable.

Maximal Margin Classifier

Write (1) in standard constrained optimisation form:

Definition (maximal margin classifier)

The maximal margin classifier solves the problem

$$\begin{aligned} &\max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad M \\ \text{subject to} \quad & \|w\| = 1 \\ & y_i(w\cdot x_i + b) \geq M, \forall i \end{aligned} \tag{2}$$

- The equivalent form of maximal margin classifier is

$$\begin{aligned} &\max_{w\in\mathbb{R}^d, b\in\mathbb{R}} \quad M \\ \text{subject to} \quad & y_i(w\cdot x_i + b)/\|w\| \geq M, \quad i = 1, 2, \ldots, n \end{aligned} \tag{3}$$

- 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 $\phi^* = \operatorname{sign} f^*$.

Exercise (XOR)

Suppose the dataset has $n = 4$ examples in $\mathbb{R}^2$ plane as follows:

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

. Find the maximal margin classifier

$f(x) = w_1 x_{(1)} + w_2 x_{(2)} + b$ *where* $x = (x_{(1)}, x_{(2)}) \in \mathbb{R}^2$

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

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

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 $f(x) = (-x_1 + x_2 \pm 1)/\sqrt{2}$. If $M = -b$, then $w_1 = w_2 = \pm b$ and the solution is $f(x) = (-x_1 - x_2 \pm 1)/\sqrt{2}$

Cover theorem: increasing dimension improves linear separability

Cover's theorem:

> *"pattern-classification problem cast in a high dimensional space non-linearly is more likely to be linearly separable than in a low-dimensional space"*

Kernel trick:

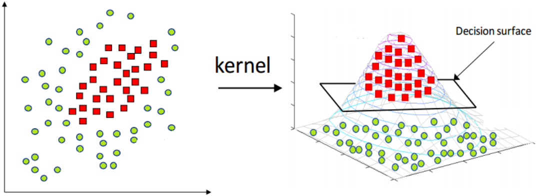


image source

Recall the maximal margin classifier (3) Since we can scale w, b by a **positive** factor arbitrarily, we can assume $M > 0$ and use the normalization $M\|w\| = 1$ *if the dataset is linearly separable*, instead of using the normalization $\|w\| = 1$. Then

$$\begin{aligned} &\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \\ &\text{subject to} \quad \boxed{y_i(w\cdot x_i + b) \geq 1, \forall i} \end{aligned} \tag{4}$$

- Now there is NO admissible solution if the dataset is not linearly separable, in contrast to (3) and (2).
- The problem (4) is the standard quadratic programming problem ☺.
- The margin now is $M = \frac{1}{\|w\|}$.
- The support vectors are those on the two hyperplanes

  $$\boxed{w\cdot x + b = \pm 1},$$

  i.e., the inequality constraints at these support vectors actually are equalities.

Support Vector Classifier

soft margin and slack variable for nonseparable dataset

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)

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \frac{1}{2} \|w\|^2 \tag{5}$$

$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i \tag{6}$$

$$\xi_i \geq 0, \forall i \tag{7}$$

$$\sum_{i=1}^{n} \xi_i \leq s \tag{8}$$

where the constant $s > 0$ is a tuning parameter.

This is in the standard form of quadratic programming.

Remarks on relaxation budget

- hyperparameter s controls the budget of relaxation:
  - $s = 0 \iff \xi_i \equiv 0$, maximal margin classifier becomes (4)(for linearly separable case) and does not allow violation of the margin.
  - If $s = +\infty$, any w and b are admissible, and then the optimal $w^* = 0$, b arbitrary: huge bias, low variance
- The budget $\sum_i \xi_i < s$ can be rewritten as the penalty form with the parameter $C > 0$

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \tag{9}$$

$$\text{subject to} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \forall i \tag{10}$$

$$\xi_i \geq 0, \forall i \tag{11}$$

Totally, $d + 1 + n$ unknowns.

Interpretation of slack variables at optimality

There are two constraints:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \text{and } \xi_i \geq 0$$

At the optimal values, some constraints may be active, some others may be inactive.
With the abuse of language, we define "margin" as the set between two hyperplanes $\mathcal{M} := \{x : |w \cdot x + b| \leq 1\}$

- $\xi_i = 0$, then x_i is correctly classified and even better that it is not inside of the margin $\mathcal{M}$. Furthermore, if additionally $y_i(w \cdot x_i + b) = 1$, then x_i is a support vector, sitting on the boundary of margin.
- $\xi_i > 0$, then $y_i(w \cdot x_i + b) = 1 - \xi_i$ and is less than 1, x_i violates the margin boundary and enters $\mathcal{M}$;
    - $0 < \xi < 1$: x_i is still correctly classified but too close to the decision boundary.
    - $\xi_i > 1$: $y_i(w \cdot x_i + b)$ is negative so, x_i is misclassified.

1

[1] The theory of constrained optimization such as linear programming theory is helpful in understanding this part.

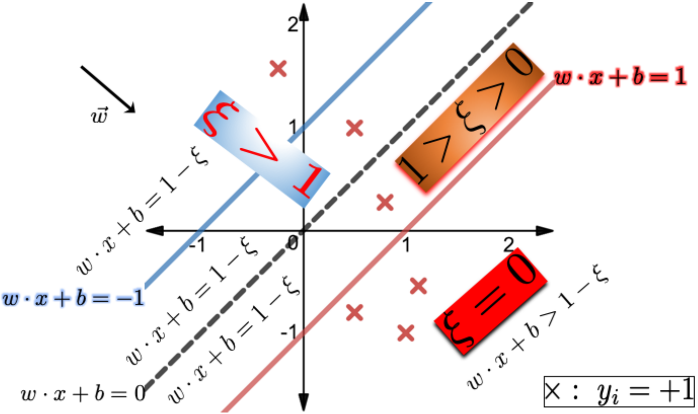


Figure: The slack variables for the data point in class $+1$, markers("$\times$"). $x \in \mathbb{R}^2$.

The larger value of ξ_i, the further the point x_i away from the correct domain. This justifies the penalty of $\sum_i \xi_i$.

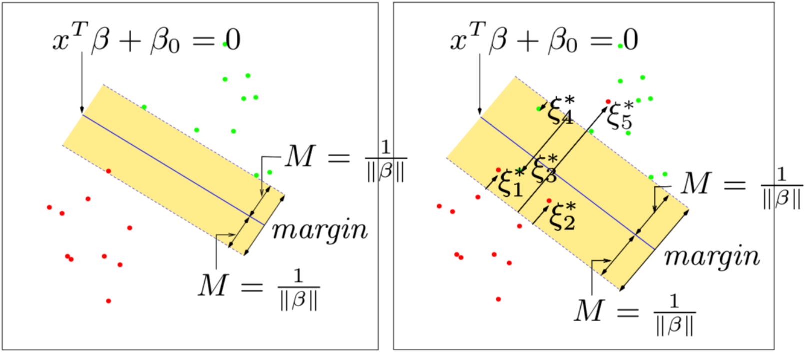


FIGURE 12.1. *Support vector classifiers. The left*

From [ESL]: here $\xi_i^* := M\xi_i$

Exercise: discuss the range of $\xi_i^* = M\xi_i$ in the right figure.

Interpretation of Support vectors

The support vectors are those x_i such that $y_i(w \cdot x_i + b) = 1 - \xi_i$. This equation of w, b can easily solved if these support vectors as well as ξ_i are known. In fact the solution of SVC is

$$w^* = \sum_{i \in S} \hat{\alpha} y_i x_i$$

where $S = \{i : y_i(w \cdot x_i + b) = 1 - \xi_i\}$. Refer to Section 12.2.1 in [ESL].

- Complexity of the classifier is characterized by the number of support vectors rather than the dimensionality.
- The solution is insensitive to the outliers (the data points significantly far away from the decision boundary).

SVM with kernel tricks: nonlinear decision boundaries

The key idea of extending linear SVC, and many other linear procedures, to nonlinear is to:

- Enlarge the predictor space using basis expansion functions $h_1(x), \ldots, h_M(x)$
- Construct a linear separating hyperplane $f(x) = w \cdot h(x) + b$ in the enlarged space for better training performance
- The linear separating hyperplane in the enlarged space can be translated into a nonlinear separating hyperplane in the original space.

The procedure is to replace x in SVC by $h(x)$ in SVM. We do not discuss the details further: this is a general principle.

NEXT:
Rewrite the Constraint Optimization form of SVC into the classic form
Loss function + Penalty

Note that two constraints in SVC $y_i f(x_i) \geq 1 - \xi_i$ and $\xi_i \geq 0$ together are equivalent to

$$\xi_i \geq \max\{0, 1 - y_i f(x_i)\} =: (1 - y_i f(x_i))_+.$$

SVM= hinge Loss + L_2-Regularization

Then the SVC in (9) is equivalent to

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

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

which is equivalent to

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2C}\|w\|^2 + \sum_i (1 - y_i(w\cdot x_i + b))_+ \tag{12}$$

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

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

The population risk is

$$\mathcal{E}(f) = \mathbb{E}\,\ell_{hinge}(Y, f(X))$$

The penalty for the roughness of f : $R(f) = \|f\|$.
When $f(x) = w \cdot x + b$,

$$\min_{w,b} \sum_{i=1}^{n} \ell_{hinge}(y_i, f(x_i)) + \frac{\lambda}{2} \|w\|^2 \tag{13}$$

- $\lambda = 1/C$: the budget for relaxation.
- $\lambda > 0$: the penalty on the margin-relevant variable $\|w\|$.
- The soft margin performs like regularization; so the SVM is usually good at generalization.

logistic regression : binomial deviance Loss without Regularization

Recall the logistic regression solves

$$\min_f \mathbb{E}\,\ell_{bd}(Y, f(X)) \approx \frac{1}{n}\sum_{i=1}^{n} \ell_{bd}(y_i, f(x_i)) \tag{14}$$

where the binomial deviance loss

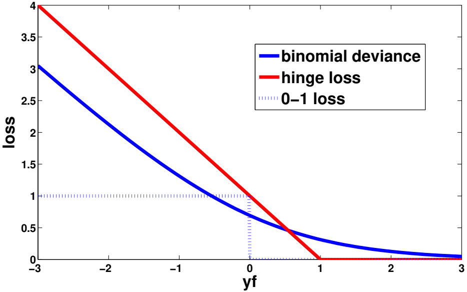
$$\ell_{bd}(y, f) = \log(1 + e^{-yf}), \quad y \in \{-1, 1\}\,.$$

We know the optimal $f^*_{bd}(x) = \text{logit}(h) = \log \frac{h}{1-h}$ where $h(x) = \mathbb{P}(Y = +1|X = x)$. The classifier is

$$x \to \text{sign}(f^*_{bd}(x))$$

This is equivalent to the Bayes classifier with the 0-1 loss $\mathbb{E}\,\ell_{01}(Y, G(X))$: $G^*(x) = \text{sign}(h(x) - 0.5) = \text{sign}\, f^*_{bd}(x)$

Then we have three loss functions ℓ_{bd}, ℓ_{hinge} and ℓ_{01}. They are functions in term of the product $yf(x)$. More choice of loss functions are in Table 12.1 [ESL].



discussion: What differences ? Computational issues ? Which data examples feel the "gradient" force?

Exercise

Consider the minimization problem

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

for the $\{\pm 1\}$*-encoded binary classification problem. Show that the optimal* f^*_{hinge} *is sign*$(h(x) - 0.5)$ *where* $h(x) = \mathbb{P}(Y = +1 | X = x)$*.*

Recall a similar exercise for the binomial deviance loss.

- Note sign$(f^*_{hinge}) = f^*_{hinge}$ since f^*_{hinge} only takes value ± 1.
- f^*_{bd} is continuous, but f^*_{hinge} is a step-type function.

$$
\begin{aligned}
&\mathbb{E}\,\ell_{hinge}(Y, f(X)) \\
=&\pi_+\,\mathbb{E}_{X|Y=+1}\,\ell_{hinge}(+1, f(X)) + \pi_-\,\mathbb{E}_{X|Y=-1}\,\ell_{hinge}(+1, f(X)) \\
=&\pi_+ \int_{\mathcal{X}} I(1 - f(x) > 0)\,(1 - f(x))\,\rho_+(x)\mathrm{d}x \\
&\quad + \pi_- \int_{\mathcal{X}} I(1 + f(x) > 0)\,(1 + f(x))\,\rho_-(x)\mathrm{d}x
\end{aligned}
$$

Consider the domain $\Omega_+ \subset \mathcal{X}$ where $f(x) > 1$, then the integration over Ω_+ is $\pi_- \int_{\Omega_+}(1 + f(x))\rho_-(x)\mathrm{d}x$: by decreasing the value of f on this domain Ω_+ to the minimal possible value 1, one has a smaller loss. So, for f^* to be optimal, Ω_+ must be empty. For the same reason for Ω_- case, we deduce that $f^*(x) \in [-1, 1]$ almost everywhere. Then we only consider to minimize within this bounded function class $\{f : |f(x)| \leq 1\}$ $\pi_+ \int_{\mathcal{X}}(1 - f(x))\,\rho_+(x)\mathrm{d}x + \pi_- \int_{\mathcal{X}}(1 + f(x))\,\rho_-(x)\mathrm{d}x =$ $\int_{\mathcal{X}} f(x)\,(\pi_-\rho_-(x) - \pi_+\rho_+(x))\,\mathrm{d}x + 1$. So if $\pi_-\rho_-(x) < \pi_+\rho_+(x)$, the minimizer is $f^*(x) = 1$; otherwise $f^*(x) = -1$. Equivalently, $f^*(x) = \mathrm{sign}(\pi_+\rho_+(x) - \pi_+\rho_+(x)) = \mathrm{sign}(h(x) - 0.5)$ since $h(x) = \frac{\pi_+\rho_+(x)}{\pi_+\rho_+(x)+\pi_+\rho_+(x)}$

Regularization in SVM smoothes f^*_{hinge}

- The objective function with regularization

  $$\inf_{f\in\mathcal{H}} \mathbb{E}\,\ell_{hinge}(Y, f(X)) + \lambda R(f)$$

  gives a smooth function approximation f_λ to the step function f^*_{hinge}, and the classifier is

  $$x \to \mathsf{sign}(f_\lambda)$$

  places back f_λ to step function by thresholding.
- For SVC, the hypothesis space $\mathcal{H}$ is linear function space, and SVC use a linear function to approximate a step function with jumps. The maximal margin $\|w\|$ has the meaning of penalty.

topics not touched here for SVM

- kernel trick, RKHS
- optimization theory and methods for SVM