CS:E4830 Kernel Methods in Machine Learning

Lecture 8 : Kernel Support Vector Machines

Rohit Babbar

6th March, 2019

Hinge Loss Function and SVM

- Hinge loss is a function $\mathbb{R} \mapsto \mathbb{R}_+$:

$$\ell_{hinge}(u) = max(1-u, 0) = \begin{cases} 0 & \text{if } u \geq 1 \\ 1-u & \text{otherwise} \end{cases}$$

- SVM solves the following optimization problem :

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i f(x_i)) + \lambda ||f||_{\mathcal{H}}^2$$

Hinge Loss and Others

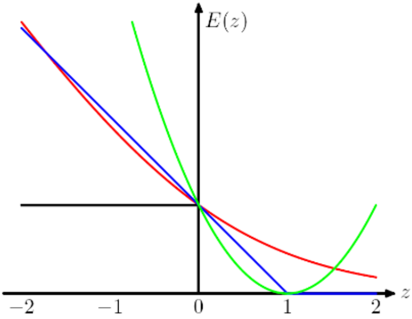


Figure: $z = yf(x)$ in the above graph

Convex Upper Bounds on 0-1 loss

- Hinge Loss (in blue) is given by $\max(1 - yf(x), 0)$
- Logistic Loss is given by $\frac{1}{\log 2} \log(1 + \exp(-yf(x)))$ (re-scaled version compared to previous lecture)

Hinge Loss and Others

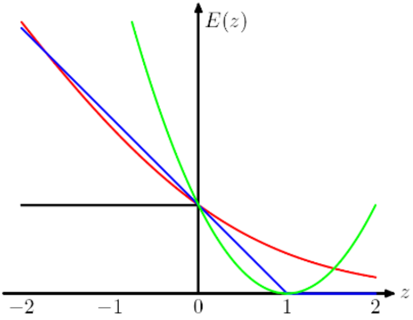


Figure: $z = yf(x)$ in the above graph

Convex Upper Bounds on 0-1 loss

- Hinge Loss (in blue) is given by $\max(1 - yf(x), 0)$
- Logistic Loss is given by $\frac{1}{\log 2} \log(1 + \exp(-yf(x)))$ (re-scaled version compared to previous lecture)

From Representer Theorem

- For the following optimization

$$f_{\mathcal{H}} := \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i, f(x_i)) + \lambda\theta(||f||_{\mathcal{H}}^2)$$

where $\ell_{hinge}(.,.)$ is the hinge loss function and $\theta : [0, \infty) \mapsto \mathbb{R}$ is non-decreasing function, and $\mathcal{H}$ is an RKHS

- Even though the above problem is potentially an infinite dimensional optimization problem, **Representer Theorem** states its solution can be expressed in the following form

$$f(.) = \sum_{j=1}^{n} \alpha_i k(., x_j)$$

where $\alpha_i \in \mathbb{R}$, i.e. it is linear combination of kernel evaluations at training points

SVM Problem Refomulation

- Using Representer theorem, the problem can be reformulated as

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i [K\alpha]_i) + \lambda \alpha^T K \alpha \right\}$$

SVM Problem Refomulation

- Using Representer theorem, the problem can be reformulated as

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i [K\alpha]_i) + \lambda \alpha^T K \alpha \right\}$$

- The above optimization problem is convex (Why?)

SVM Problem Refomulation

- Using Representer theorem, the problem can be reformulated as

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i[K\alpha]_i) + \lambda \alpha^T K \alpha \right\}$$

- The above optimization problem is convex (Why?)
- However, it is non-smooth optimization problem (Why?)

SVM Problem Refomulation

- Using Representer theorem, the problem can be reformulated as

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i[K\alpha]_i) + \lambda \alpha^T K \alpha \right\}$$

- The above optimization problem is convex (Why?)
- However, it is non-smooth optimization problem (Why?)

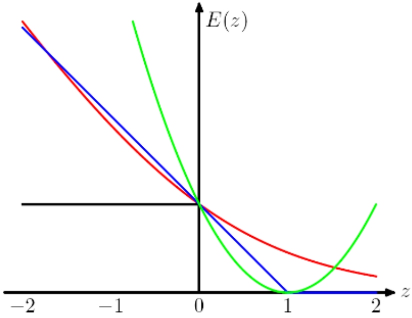


Figure: $z = yf(x)$ in the above graph

Another Equivalent Reformulation

- The optimization problem on the previous slide is equivalent (even though not immediately obvious) to the following, if we re-write it in terms of slack variables $\xi_i \in \mathbb{R}$ for $i = 1, \ldots, n$

$$\min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha \right\} \text{such that } \xi_i \geq \ell_{hinge}(y_i [K\alpha]_i)$$

Another Equivalent Reformulation

- The optimization problem on the previous slide is equivalent (even though not immediately obvious) to the following, if we re-write it in terms of slack variables $\xi_i \in \mathbb{R}$ for $i = 1, \ldots, n$

$$\min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha \right\} \text{such that } \xi_i \geq \ell_{hinge}(y_i [K\alpha]_i)$$

- In the above formulation, the objective is smooth but not the constraints

Another Equivalent Reformulation

- The optimization problem on the previous slide is equivalent (even though not immediately obvious) to the following, if we re-write it in terms of slack variables $\xi_i \in \mathbb{R}$ for $i = 1, \ldots, n$

$$\min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha \right\} \text{such that } \xi_i \geq \ell_{hinge}(y_i [K\alpha]_i)$$

- In the above formulation, the objective is smooth but not the constraints
- Recall the definition of hinge loss from first slide

$$\ell_{hinge}(u) = max(1 - u, 0) \iff \begin{cases} 0 & \text{if } u \geq 1 \\ 1 - u & \text{otherwise} \end{cases}$$

- Using above, the n constraints ($\xi_i \geq \ell_{hinge}(y_i[K\alpha]_i)$) can be replaced by $2n$ constraints to make the problem smooth as follows :

$$\xi_i \geq \ell_{hinge}(y_i[K\alpha]_i) \iff \begin{cases} \xi_i \geq & 1 - y_i[K\alpha]_i \\ \xi_i \geq & 0 \end{cases}$$

Putting Things Together

- To summarize, the SVM solution is given by

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x, x_i)$$

where $\hat{\alpha}$ is the solution to the following :

SVM Primal Formulation

$$\min_{\alpha \in \mathbb{R}^n, \xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha \right\}$$

such that

$$\begin{cases} 1 - y_i[K\alpha]_i - \xi_i \leq 0 & \text{for } i = 1, \ldots, n \\ -\xi_i \leq 0 & \text{for } i = 1, \ldots, n \end{cases}$$

Lagrangian

- The Lagrangian of the problem is :

$$L(\alpha, \xi, \mu, \nu) = \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha + \sum_{i=1}^{n} \mu_i [1 - y_i [K\alpha]_i - \xi_i] - \sum_{i=1}^{n} \nu_i \xi_i$$

- Note that constraints have moved to the Lagrangian.

Lagrangian wrt α

- The lagrangian of the problem is :

$$L(\alpha, \xi, \mu, \nu) = \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\alpha^T K\alpha + \sum_{i=1}^{n}\mu_i[1 - y_i[K\alpha]_i - \xi_i] - \sum_{i=1}^{n}\nu_i\xi_i$$

Lagrangian wrt α

- $L(\alpha, \xi, \mu, \nu)$ is a convex quadratic function in α. To find the optimal value, set the gradient to $\mathbf{0}$ (the zero vector) :

$$\nabla_\alpha L = \mathbf{0}$$

- The optimal solution α^* is given by

$$\alpha_i^* = \frac{y_i\mu_i}{2\lambda}$$

Lagrangian wrt ξ

- The lagrangian of the problem is :

$$L(\alpha, \xi, \mu, \nu) = \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\alpha^T K\alpha + \sum_{i=1}^{n}\mu_i[1 - y_i[K\alpha]_i - \xi_i] - \sum_{i=1}^{n}\nu_i\xi_i$$

Lagrangian wrt ξ

- $L(\alpha, \xi, \mu, \nu)$ is a linear function in ξ.
- Its minimum value is $-\infty$, except when it is constant,

$$\nabla_\xi L = \frac{\mathbf{1}}{n} - \mu - \nu = \mathbf{0}$$

equivalently,

$$\frac{\mathbf{1}}{n} = \mu + \nu$$

Lagrange Dual Function and Dual Problem

Lagrange Dual Function

- The Lagrange dual function as obtained by substituting the optimal values (as obtained in previous two slides) is given by :

$$q(\mu,\nu) = \min_{\alpha\in\mathbb{R}^n,\xi\in\mathbb{R}^n} L(\alpha,\xi,\mu,\nu)$$
$$= \begin{cases} \sum_{i=1}^{n}\mu_i - \frac{1}{4\lambda}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \mu_i \mu_j K(x_i,x_j) & \text{if } \mu+\nu=\frac{1}{n} \\ -\infty & \text{otherwise} \end{cases}$$

Lagrange Dual Problem

- The Lagrange dual problem is

$$\max q(\nu,\mu) \text{ such that } \mu \geq 0, \nu \geq 0$$

Closer Look At The Dual Problem

- The Lagrange dual problem is

$$\max q(\nu, \mu) \text{ such that } \mu \geq 0, \nu \geq 0$$

- If $0 \leq \mu_i \leq 1/n$ for all i, then the dual function takes finite values. Also, the value of ν_i is fixed at $\nu_i = 1/n - \mu_i$ in this case.
- The dual problem is therefore given by

$$\max_{\mathbf{0} \leq \mu \leq \mathbf{1}/n} \sum_{i=1}^{n} \mu_i - \frac{1}{4\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \mu_i \mu_j K(x_i, x_j)$$

Rewriting in terms of Primal Variables

Dual problem (from previous slide)

$$\max_{\mathbf{0} \leq \mu \leq \mathbf{1}/n} \sum_{i=1}^{n} \mu_i - \frac{1}{4\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \mu_i \mu_j K(x_i, x_j)$$

Since the primal variable α and the dual variable μ are related by $\alpha_i = \frac{\mu_i y_i}{2\lambda}$, it can be written in the form of primal variables as follows

writing in terms of primal variable α

$$\max_{\alpha \in \mathbb{R}^n} 2 \sum_{i=1}^{n} \alpha_i y_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$

such that

$$0 \leq y_i \alpha_i \leq \frac{1}{2\lambda n} \text{ for } i = 1, \ldots, n$$

Complementarity conditions at the optimum

- These are given by the product of the dual variables and the corresponding constraint as follows :

$$\mu_i[y_i f(x_i) + \xi_i - 1] = 0$$
$$\nu_i \xi_i = 0$$

- In terms of the primal variable α, it is given by

$$\alpha_i[y_i f(x_i) + \xi_i - 1] = 0$$
$$\left(\alpha_i - \frac{y_i}{2\lambda n}\right) \xi_i = 0$$

Interpreting the Complementarity Conditions

Complementarity Conditions

$$\alpha_i[y_i f(x_i) + \xi_i - 1] = 0$$
$$\left(\alpha_i - \frac{y_i}{2\lambda n}\right)\xi_i = 0$$

- If $\alpha_i = 0$, then the second constraint is active : $\xi_i = 0$. This implies $y_i f(x_i) \geq 1$

Interpreting the Complementarity Conditions

Complementarity Conditions

$$\alpha_i[y_i f(x_i) + \xi_i - 1] = 0$$
$$\left(\alpha_i - \frac{y_i}{2\lambda n}\right)\xi_i = 0$$

- If $\alpha_i = 0$, then the second constraint is active : $\xi_i = 0$. This implies $y_i f(x_i) \geq 1$
- If $0 < y_i \alpha_i < \frac{1}{2\lambda n}$, then both the constraints are active, i.e., $\xi_i = 0$ and $y_i f(x_i) + \xi_i - 1 = 0$. This leads to $y_i f(x_i) = 1$.

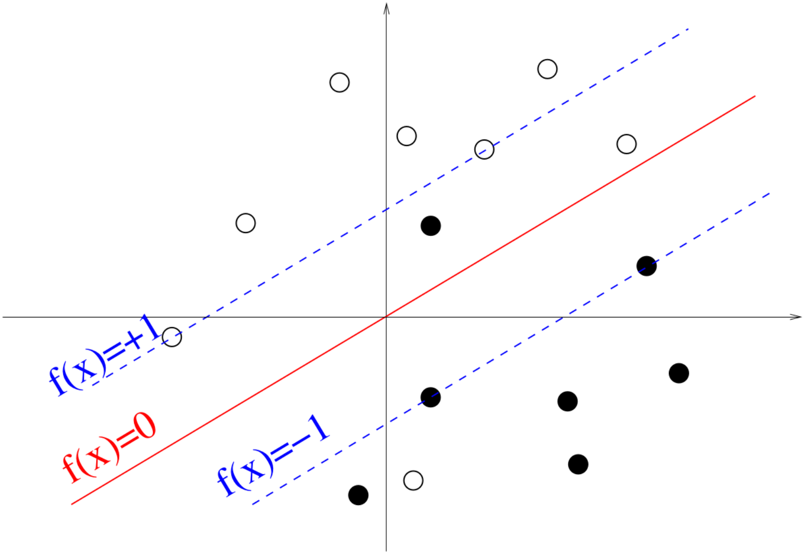
Interpreting the Complementarity Conditions

Complementarity Conditions

$$\alpha_i[y_i f(x_i) + \xi_i - 1] = 0$$
$$\left(\alpha_i - \frac{y_i}{2\lambda n}\right)\xi_i = 0$$

- If $\alpha_i = 0$, then the second constraint is active : $\xi_i = 0$. This implies $y_i f(x_i) \geq 1$
- If $0 < y_i\alpha_i < \frac{1}{2\lambda n}$, then both the constraints are active, i.e., $\xi_i = 0$ and $y_i f(x_i) + \xi_i - 1 = 0$. This leads to $y_i f(x_i) = 1$.
- If $\alpha_i = \frac{y_i}{2\lambda n}$, then the second constraint is not active ($\xi_i \geq 0$) but the first one is active : $y_i f(x_i) + \xi_i = 1$. This implies that $y_i f(x_i) \leq 1$.

Decision Hyperplanes



Pictorial Depiction for α values

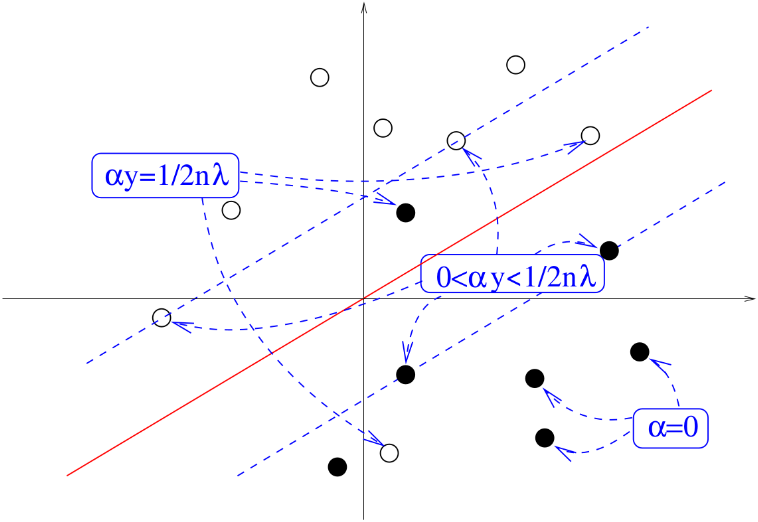


Figure: Picture : Julien Mairal

Support Vectors

- From Representer theorem, the function evaluation at any $x \in \mathcal{X}$ (the input space) is given by

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x) = \sum_{i \in SV} \alpha_i k(x_i, x)$$

  where SV is the set of support vectors i.e. those training points for which $\alpha_i \neq 0$
- Hence the name Support Vector Machines
- The above sparsity of $\boldsymbol{\alpha} \in \mathbb{R}^n$ can be used for
  - Faster prediction since once needs to go over only the support vectors

Another variant - C-SVM

Sometimes, instead of the regularization parameter λ, the SVM problem is written in the following form :

$$\min_{f \in \mathcal{H}} \left\{ C \sum_{i=1}^{n} (\ell_{hinge}(y_i[K\alpha]_i))^2 + \frac{1}{2}||f||_{\mathcal{H}}^2 \right\}$$

Another variant - C-SVM

Sometimes, instead of the regularization parameter λ, the SVM problem is written in the following form :

$$\min_{f \in \mathcal{H}} \left\{ C \sum_{i=1}^{n} (\ell_{hinge}(y_i[K\alpha]_i))^2 + \frac{1}{2}||f||_{\mathcal{H}}^2 \right\}$$

- This is equivalent to the original formulation on the first slide($\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i f(x_i)) + \lambda ||f||_{\mathcal{H}}^2$) with $C = \frac{1}{2n\lambda}$

Another variant - C-SVM

Sometimes, instead of the regularization parameter λ, the SVM problem is written in the following form :

$$\min_{f \in \mathcal{H}} \left\{ C \sum_{i=1}^{n} (\ell_{hinge}(y_i [K\alpha]_i))^2 + \frac{1}{2} ||f||_{\mathcal{H}}^2 \right\}$$

- This is equivalent to the original formulation on the first slide($\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell_{hinge}(y_i f(x_i)) + \lambda ||f||_{\mathcal{H}}^2$) with $C = \frac{1}{2n\lambda}$
- Using the Lagrangian formulation, the dual can be written as

$$\max_{\alpha \in \mathbb{R}^n} 2 \sum_{i=1}^{n} \alpha_i y_i - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$

such that

$$0 \leq y_i \alpha_i \leq C \text{ for } i = 1, \ldots, n \text{ (also called box constraints)}$$

- Most of the material for this lecture is based on a similar course by Julien Mairal's at ENS Paris
- Further details (with somewhat different notation) on SVMs - JST & Christianini book, Chapter 7