CS5487 Problem Set 9

Support Vector Machines

Antoni Chan
Department of Computer Science
City University of Hong Kong

Hyperplanes

Problem 9.1 Margin

Let $f(x) = w^T x + b$ and consider the hyperplane $f(x) = 0$.

(a) Show that the Euclidean distance from a point x_a to the hyperplane is $\frac{|f(x_a)|}{\|w\|}$ by minimizing $\|x - x_a\|^2$ subject to $f(x) = 0$.

(b) Show that the distance from the origin to the hyperplane is $\frac{|b|}{\|w\|}$.

(c) Show that the projection of x_a onto the hyperplane is

$$x_p = x_a - \frac{f(x_a)}{\|w\|^2} w. \tag{9.1}$$

....

SVMs

Problem 9.2 SVM dual problem

Consider the SVM problem for linear separable data $\{X, y\}$,

$$\min_{w,b} \frac{1}{2} \|w\|^2 \text{ s.t. } y_i(w^T x_i + b) \geq 1, \quad \forall i. \tag{9.2}$$

We will derive the dual formulation of the SVM.

(a) Show that the Lagrangian of (9.2) is

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + b) - 1), \tag{9.3}$$

where $\alpha_i \geq 0$ is a Lagrange multiplier for each i.

(b) Show that the minimum of $L(w, b, \alpha)$ w.r.t. $\{w, b\}$ satisfies

$$w^* = \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{9.4}$$

(c) Use (9.4) on the Lagrangian to obtain the dual function,

$$L(\alpha) = \min_{w,b} L(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j, \tag{9.5}$$

yielding the SVM dual problem, $\max_\alpha L(\alpha)$,

$$\begin{aligned} \max_{\alpha} \ & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \ & \sum_{i=1}^{n} \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \ \forall i. \end{aligned} \tag{9.6}$$

....

Problem 9.3 Calculating the bias term

Solving the dual SVM problem yields only the hyperplane parameters w. In this problem, we consider two ways to calculate the bias term b. Recall that the set of support vectors are those training points with non-zero Lagrange multipliers,

$$SV = \{i | \alpha_i > 0\}. \tag{9.7}$$

Since $\alpha_i > 0$ for points in SV, by the KKT conditions, the equality constraint must be "active" for the support vector x_i, i.e.,

$$y_i(w^T x_i + b) = 1. \tag{9.8}$$

(a) Using the active constraints, show that b can be calculated as

$$b^* = \frac{1}{|SV|}\sum_{i \in SV} (y_i - w^T x_i) \tag{9.9}$$

(b) Another method for calculating the bias is to select one support vector on each side of the hyperplan. Given points x^+ and x^-, which are on the positive and negative margins, show that

$$b^* = -\frac{1}{2} w^T (x^+ + x^-) \tag{9.10}$$

(c) Show that the method in (b) is a special case of (a) when there are only two support vectors.

....

Problem 9.4 Soft-margin SVM (1-norm penalty)

Consider the soft-margin SVM we saw in lecture,

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\ &\qquad \xi_i \geq 0, \quad \forall i \end{aligned} \tag{9.11}$$

where ξ_i is the slack variable that allows the ith point to violate the margin. The new term in the objective function penalizes large slack variables (with parameter C). Since the penalty on the slack variable is their sum and the slack is non-negative, then the penalty function is the 1-norm of the slack variables. We will derive the dual formulation of this SVM. (This SVM is sometimes called *C-SVM* or *1-norm SVM*. It is also the most popular one used.)

(a) Introduce Lagrange multipliers $\alpha_i \geq 0$ for the margin constraint, and $r_i \geq 0$ for the non-negative slack constraint, show that the Lagrangian is

$$L(w, b, \xi, \alpha, r) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} r_i\xi_i, \tag{9.12}$$

(b) Show that the minimum of $L(w, b, \xi, \alpha, r)$ w.r.t. $\{w, b, \xi\}$ satisifies

$$w^* = \sum_{i=1}^{n}\alpha_i y_i x_i, \qquad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad r_i = C - \alpha_i. \tag{9.13}$$

(c) Use (9.13) on the Lagrangian to obtain the dual function,

$$L(\alpha) = \min_{w,b,\xi} L(w, b, \xi, \alpha, r) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j. \tag{9.14}$$

(d) Show that the two non-negative constraints on the Lagrange multipliers $\{\alpha_i, r_i\}$ and the equality constraint from (9.13),

$$\alpha_i \geq 0, \tag{9.15}$$
$$r_i \geq 0, \tag{9.16}$$
$$r_i = C - \alpha_i \tag{9.17}$$

are equivalent to

$$0 \leq \alpha_i \leq C. \tag{9.18}$$

Hence, the SVM dual problem is

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &\text{s.t. } \sum_{i=1}^{n}\alpha_i y_i = 0, \\ &\qquad 0 \leq \alpha_i \leq C, \quad \forall i. \end{aligned} \tag{9.19}$$

(e) Use the KKT conditions to show that only one of these conditions hold for each point x_i:

   i. $\xi_i = 0$ and $0 < \alpha_i < C$, and the point is on the margin.
   ii. $\xi_i > 0$ and $\alpha_i = C$, and the point is an outlier (violates the margin).
   iii. $\xi_i = 0$ and $\alpha_i = 0$, and the point is correctly classified.

....

Problem 9.5 Soft-margin SVM risk function

Consider the soft-margin SVM in Problem 9.4. In this problem we will interpret the soft-margin SVM as regularized risk minimization.

(a) Show that the slack variables must satisfy,

$$\xi_i \geq \max(0, 1 - y_i(w^T x_i + b)). \tag{9.20}$$

Given the above, at the optimum of the SVM primal problem, what is the expression for ξ_i?

(b) Show that the C-SVM primal problem is equivalent to

$$\min_{w,b} \sum_{i=1}^{n} \max(0, 1 - y_i(w^T x_i + b)) + \lambda \|w\|^2. \tag{9.21}$$

The first term is the empirical risk, where the loss function is

$$L_{SVM}(z_i) = \max(0, 1 - z_i). \tag{9.22}$$

This is called the "hinge loss". The second term is a regularization term on w to control its complexity. Hence, SVM is optimizing a regularized risk.

(c) Plot the SVM loss function along with the other loss functions from Problem 8.5. Give an intuitive explanation about why the SVM loss function is good (and possibly better than others).

....

Problem 9.6 Soft-margin SVM (2-norm penalty)

Consider the soft-margin SVM problem using a 2-norm penalty on the slack variables,

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i^2 \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\ &\qquad \xi_i \geq 0, \quad \forall i \end{aligned} \tag{9.23}$$

where ξ_i is the slack variable that allows the ith point to violate the margin. We will derive the dual of this SVM.

(a) Show that the non-negative constraint on ξ_i is redundant, and hence can be dropped. Hint: show that if $\xi_i < 0$ and the margin constraint is satisfied, then $\xi_i = 0$ is also a solution with lower cost.

(b) Introduce Lagrange multipliers α_i for the margin constraint, show that the Lagrangian is

$$L(w,b,\xi,\alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i). \tag{9.24}$$

(c) Show that the minimum of $L(w,b,\xi,\alpha)$ w.r.t. $\{w,b,\xi\}$ satisifies

$$w^* = \sum_{i=1}^{n}\alpha_i y_i x_i, \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \quad \xi_i = \frac{\alpha_i}{C}. \tag{9.25}$$

(d) Use (9.25) on the Lagrangian to obtain the dual function,

$$L(\alpha) = \min_{w,b,\xi} L(w,b,\xi,\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i^T x_j + \frac{1}{C}\delta_{ij}), \tag{9.26}$$

where $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. Hence, the SVM dual problem is

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i^T x_j + \frac{1}{C}\delta_{ij}) \\ &\text{s.t.} \sum_{i=1}^{n}\alpha_i y_i = 0, \\ &\quad\ \alpha_i \geq 0, \quad \forall i. \end{aligned} \tag{9.27}$$

(e) (9.27) is the same as the standard SVM dual, except for the extra $1/C$ term. What is the role of this term, and how does it affect the solution?

....

Problem 9.7 ν-SVM

One limitation of the soft-margin SVM using the 1-norm penalty (Problem 9.4) is that there is no intuition for what the parameter C means and it can therefore be difficult to find good values for it in practice. In this problem we consider a slightly different, but more intuitive formulation, based on the solution of the following problem, with ν as the parameter

$$\begin{aligned} &\min_{w,\xi,\rho,b} \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } y_i(w^T x_i + b) \geq \rho - \xi_i, \quad \forall i \\ &\qquad \xi_i \geq 0, \quad \forall i, \\ &\qquad \rho \geq 0. \end{aligned} \tag{9.28}$$

(a) Derive the dual problem and the resulting decision function.

(b) Given the dual solution how would you determine the values of b and ρ?

(c) Define the fraction of margin errors as

$$\epsilon_\rho = \frac{1}{n}|\{i|y_i f(x_i) < \rho\}| \tag{9.29}$$

and suppose that we solve the optimization problem on a dataset with the result that $\rho > 0$. Show that

(a) ν is an upper bound on ϵ_ρ.

(b) ν is a lower bound on the fraction of vectors that are support vectors.

(d) Show that if the solution of the second problem leads to $\rho > 0$, then the first problem with C set a priori to $\frac{1}{\rho}$ leads to the same decision function.

....

Other SVMs

Problem 9.8 Adaptive SVM

In this problem, we will consider an adaptive SVM. Suppose we have used a dataset $\mathcal{D}_0$ to learn a linear classifier function $f_0(x) = w_0^T x$ with decision rule $y = \text{sign}(f_0(x))$. Since we have the classifier, we then threw away the data $\mathcal{D}_0$. Now, suppose we receive a new set of data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, and now wish to update our classifier function $f_0(x)$. To do this, we will add a "delta function" $\Delta f(x) = w^T x$ to adapt our original classifier,

$$f(x) = f_0(x) + \Delta f(x) = f_0(x) + w^T x, \tag{9.30}$$

where w is the parameter vector of the delta function.

Let's consider the case when $\mathcal{D}$ is linearly separable. We wish to maximize the margin between the updated classifier and the new training set, which yields the adaptive-SVM objective function is

$$\begin{aligned} &\min_w \frac{1}{2}\|w\|^2 \\ &\text{s.t. } y_i f_0(x_i) + y_i w^T x_i \geq 1, \quad \forall i \end{aligned} \tag{9.31}$$

(a) Show that the Lagrangian of (9.31) is

$$L(w, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^n \alpha_i (y_i f_0(x_i) + y_i w^T x_i - 1), \tag{9.32}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

(b) Show that the minimum of $L(w, \alpha)$ w.r.t. w satisifies

$$w^* = \sum_{i=1}^n \alpha_i y_i x_i. \tag{9.33}$$

(c) Show that the dual function is

$$L(\alpha) = \sum_{i=1}^{n}(1 - y_i f_0(x))\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j. \tag{9.34}$$

Hence, the ASVM dual problem is

$$\begin{aligned}
\max_{\alpha}\ & \sum_{i=1}^{n}(1 - y_i f_0(x))\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
\text{s.t. } & 0 \leq \alpha_i, \quad \forall i.
\end{aligned} \tag{9.35}$$

(d) Compare the ASVM dual in (9.34) with the original SVM dual function? What is the interpretation of the ASVM dual (considering the original SVM dual)? What is the role of the original classifier $f_0(x)$?

Note that we can also define a "soft" version of ASVM with slack variables to handle the non-linearly separable case.

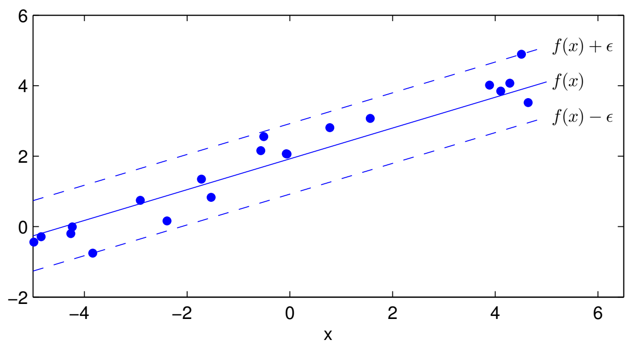
. . . .

Problem 9.9 Support Vector Regression (SVR)

In this problem, we will consider support vector regression (SVR), which applies margin principles from SVMs to regression. The goal is to learn a linear function,

$$f(x) = w^T x + b, \tag{9.36}$$

which fits a given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Suppose we form a "band" or "tube" around $f(x)$ with width ϵ (see figure below).



We can consider any training pair (x_i, y_i) that falls inside of the tube as correctly regressed, while points falling outside of the tube are errors. Assuming that ϵ is known, the SVR problem is

$$\begin{aligned}
\min_{w,b}\ & \frac{1}{2}\|w\|^2 \\
\text{s.t. } & y_i - (w^T x_i + b) \leq \epsilon, \\
& (w^T x_i + b) - y_i \leq \epsilon, \quad \forall i
\end{aligned} \tag{9.37}$$

(a) What are the roles of the inequality constraints and the objective function in (9.37)?

(b) Show that the Lagrangian of (9.37) is

$$L(w,b,\alpha,\hat{\alpha}) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i(\epsilon - y_i + (w^T x_i + b)) - \sum_{i=1}^{n}\hat{\alpha}_i(\epsilon + y_i - (w^T x_i + b), \qquad (9.38)$$

where $\alpha_i \geq 0$ are the Lagrange multiplier for the first inequality constraint, and $\hat{\alpha}_i \geq 0$ are for the second inequality constraint.

(c) Show that the minimum of $L(w,b,\alpha,\hat{\alpha})$ w.r.t. $\{w,b\}$ satisfies

$$w^* = \sum_{i=1}^{n}(\alpha_i - \hat{\alpha}_i)x_i, \quad \sum_{i=1}^{n}(\alpha_i - \hat{\alpha}_i) = 0. \qquad (9.39)$$

(d) Show that the SVR dual function is

$$L(\alpha,\hat{\alpha}) = \min_{w,b} L(w,b,\alpha,\hat{\alpha}) \qquad (9.40)$$

$$= \sum_{i=1}^{n} y_i(\alpha_i - \hat{\alpha}_i) - \epsilon\sum_{i=1}^{n}(\alpha_i + \hat{\alpha}_i) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)x_i^T x_j. \qquad (9.41)$$

Hence the SVR dual problem is

$$\begin{aligned} &\max_{\alpha,\hat{\alpha}} L(\alpha,\hat{\alpha}) \\ &\text{s.t.} \sum_{i=1}^{n}\alpha_i = \sum_{i=1}^{n}\hat{\alpha}_i, \\ &0 \leq \alpha_i, \quad 0 \leq \hat{\alpha}_i \quad \forall i. \end{aligned} \qquad (9.42)$$

(e) Use the KKT conditions to show that only one of these conditions holds for the ith data point,

i. $\alpha_i = 0$ and $\hat{\alpha}_i = 0$, and the point is inside the ϵ-tube.

ii. $\alpha_i > 0$ and $\hat{\alpha}_i = 0$, and the point is on the positive margin of the tube, i.e., $y_i = f(x_i) + \epsilon$.

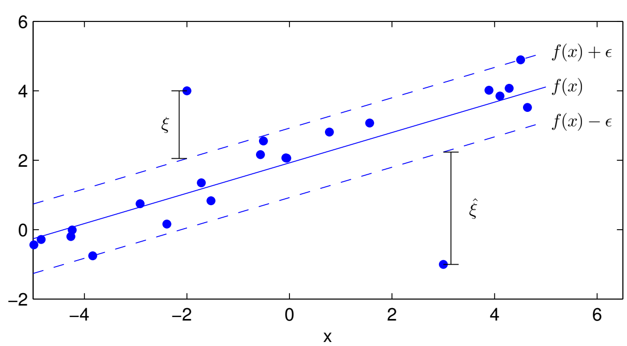
iii. $\alpha_i = 0$ and $\hat{\alpha}_i > 0$, and the point is on the negative margin of the tube, i.e., $y_i = f(x_i) - \epsilon$.

Hence, $\alpha_i\hat{\alpha}_i = 0$, i.e., the point can't both be on the positive margin and negative margin at the same time.

....

Problem 9.10 "Soft" Support Vector Regression

In this problem we consider a "soft" version of SVR. We introduce a set of slack variables $\xi_i, \hat{\xi}_i \geq 0$ to allow for some errors on either side of the tube, i.e., some points can be outside of the tube.



The amount of error is penalized linearly, similar to SVMs. The SVR problem is now:

$$\begin{aligned}
&\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) \\
&\text{s.t. } y_i - (w^T x_i + b) \le \epsilon + \xi_i, \\
&\quad (w^T x_i + b) - y_i \le \epsilon + \hat{\xi}_i, \\
&\quad \xi_i \ge 0, \quad \hat{\xi}_i \ge 0, \quad \forall i.
\end{aligned} \tag{9.43}$$

At the optimum, if the ith point is inside the positive margin of the ϵ-tube, then the slack variable $\xi_i = 0$. Otherwise, if the ith point is outside the positive margin, then ξ_i is the distance between the point and the margin, i.e., the amount of error $y_i - f(x) - \epsilon$. The same holds for $\hat{\xi}_i$ and the negative margin. In other words, the slack variables ξ_i and $\hat{\xi}_i$ contain the amount of error outside of the ϵ-tube.

(a) Let $|z|_\epsilon$ be the ϵ-insensitive loss function,

$$|z|_\epsilon = \max(0, |z| - \epsilon) = \begin{cases} 0 & |z| \le \epsilon \\ |z| - \epsilon & \text{otherwise.} \end{cases} \tag{9.44}$$

Show that the primal problem in (9.43) is equivalent to regularized regression with the ϵ-insensitive loss function,

$$\min_{w,b} \sum_{i=1}^{n} |y_i - f(x_i)|_\epsilon + \lambda \|w\|^2. \tag{9.45}$$

How does this compare with other regularized regression formulations, e.g., regularized least squares? Comment on its robustness to outliers.

(b) Show that the soft SVR dual problem is:

$$\begin{aligned}
&\max_{\alpha} \sum_{i=1}^{n} y_i(\alpha_i - \hat{\alpha}_i) - \epsilon\sum_{i=1}^{n}(\alpha_i + \hat{\alpha}_i) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)x_i^T x_j \\
&\text{s.t. } \sum_{i=1}^{n}\alpha_i = \sum_{i=1}^{n}\hat{\alpha}_i, \\
&\quad 0 \le \alpha_i \le C, \quad 0 \le \hat{\alpha}_i \le C \quad \forall i.
\end{aligned} \tag{9.46}$$

(c) Use the KKT conditions to show that only one of these conditions holds for the ith data point,

1. $\alpha_i = 0$	$\hat{\alpha}_i = 0,$	$\lvert y_i - f(x_i)\rvert < \epsilon$	inside the ϵ-tube.
2. $0 < \alpha_i < C,$	$\hat{\alpha}_i = 0,$	$y_i = f(x_i) + \epsilon$	on the positive margin of the tube.
3. $\alpha_i = 0,$	$0 < \hat{\alpha}_i < C,$	$y_i = f(x_i) - \epsilon$	on the negative margin of the tube.
4. $\alpha_i = C,$	$\hat{\alpha}_i = 0,$	$y_i > f(x_i) + \epsilon$	outside the positive margin of the tube.
5. $\alpha_i = 0,$	$\hat{\alpha}_i = C,$	$y_i < f(x_i) - \epsilon$	outside the negative margin of the tube.

....