EE 127/227AT: Optimization Models in Engineering **Section 11/12 - April 2014**

Support Vector Machines

Lecturer: Arturo Fernandez *Scribe: Arturo Fernandez*

1 Support Vector Machines Revisited

1.1 (Strictly) Separable Case: The Linear Hard-Margin Classifier

First, a note that much of the material include in these notes is based on the introduction in [1].

Consider a binary classification prediction problem. Training data are given and we denote the m observations as $\mathcal{T} = \{\mathbf{X}, \mathbf{y}\} = \{\mathbf{x}_i, y_i\}_{i=1}^m$, where $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{+1, -1\}$ (accordingly $\mathbf{X} \in \mathbb{R}^{m \times n}$). First, we will consider the case where the two classes are linearly separable. That is, by an n-dimensional decision boundary which is the result of an $n+1$-dimensional hyperplane). Furthermore, since there can exist multiple hyperplanes that split the classes, we would like to find the one with the maximal margin. The motivation for this being that the decision boundary will be prone to variations in the classes, thus minimizing expected risk, also referred to sometimes as generalization error.
Thus, we need a learning method that will optimize over the parameters $\mathbf{w} = [w_1, \ldots, w_n] \in \mathbb{R}^n$ and $b \in \mathbb{R}$ to find the optimal hyperplane, which in its general form is given by

$$d(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b = \sum_{i=1}^{n} w_i x_i + b \tag{1}$$

Furthermore, after finding this hyperplane, it is sensible that our decision rule to have the property that for an unseen data point $\mathbf{x}$:

(i) If $d(\mathbf{x}, \mathbf{w}, b) > 0$, classify $\mathbf{x}$ as class 1 (i.e. its associated $y = +1$)

(ii) If $d(\mathbf{x}, \mathbf{w}, b) < 0$, classify $\mathbf{x}$ as class 2 (i.e. its associated $y = -1$)

Thus, after finding the optimal parameter $\mathbf{w}^*, b^*$ our decision rule will take the form

$$\hat{f}(x) = \operatorname{sgn}\left(d(\mathbf{x}, \mathbf{w}^*, b^*)\right) = \operatorname{sgn}\left(\mathbf{w}^{*T} x + b^*\right), \qquad \text{where} \quad \operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases} \tag{2}$$

The $\mathbf{w}^*, b^*$ are the the optimal solution parameters to an optimization problem that we will present in the next section. If a situation arises where a point $\mathbf{x}$ evaluates to $\mathbf{w}^{*T}\mathbf{x} + b^* = 0$, then flip a coin or default to a specific class.

As mentioned earlier, the decision boundary is a result of the intersection between the decision hyperplane and the input space $\mathbb{R}^n$. The boundary itself is given by $\mathbf{w}^T x + b = 0$. Note that the decision boundary is unaffected by the scaling of the parameters $(\mathbf{w}, b) \to (\alpha\mathbf{w}, \alpha b)$. Thus, consider the idea of a *canonical hyperplane*, a concept that is tied directly to the so-called *support vectors* (SVs). A canonical hyperplane satisfies the property that

$$\min_{\mathbf{x}_i \in X} |\mathbf{w}^T \mathbf{x}_i + b| = 1 \tag{3}$$

Alternatively, a canonical hyperplane $d(\mathbf{x}, \mathbf{w}, b)$ and $f(\mathbf{x})$ are the same and equal to ± 1 for the support vectors , and $|d(\mathbf{x}, \mathbf{w}, b)| > |f(x)|$ for the other training points (non-support vectors).

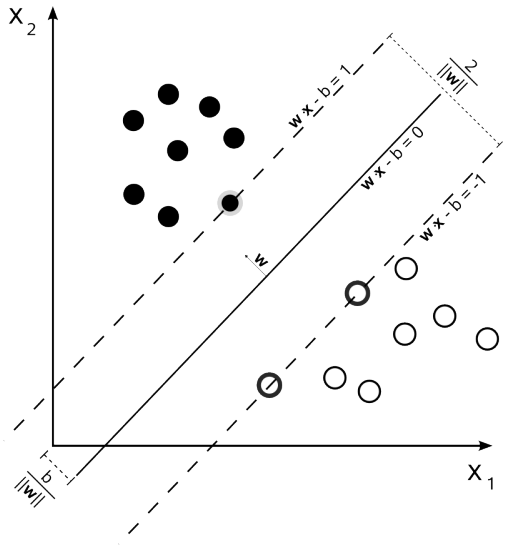

Figure 1: OCSH (Picture from Wikipedia)

1.1.1 The Optimal Canonical Separating Hyperplane

We begin the derivation of the Optimal Canonical Separating Hyperplane (OCSH) by defininining it as a *canonical* hyperplane that *separates* the data and has *maximal* margin. Thus, we nee to solve an optimization problem that maximizes the margin $M = 2/\|\mathbf{w}\|$ subject to constraints that ensure the classes are separable (Figure 1). This can be formulated as the convex optimization problem

$$\mathcal{P}1: \quad p^* := \min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad s.t. \quad y_i\left[\mathbf{w}^T\mathbf{x}_i + b\right] \geq 1, \quad i = 1, \ldots, m \tag{4}$$

where the constraints where derived from our definition of a canonical hyperplane. What about the non-separable case? Section 1.2 considers the case when the two groups are not linearly separable.

We proceed by working with the problem in its *dual space*. First, the lagrangian of (4) is

$$\mathcal{L}(\mathbf{w}, b, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{m} \lambda_i \left(1 - y_i\left[\mathbf{w}^T\mathbf{x}_i + b\right]\right) \tag{5}$$

We know by the minimax inequality that

$$p^* = \min_{\mathbf{w},b} \max_{\lambda \geq 0} \mathcal{L}(\mathbf{w}, b, \lambda) \geq \max_{\lambda \geq 0} \min_{\mathbf{w},b} \mathcal{L}(\mathbf{w}, b, \lambda) \equiv d^* \tag{6}$$

But of interest to us is when strong duality holds. Indeed, it does hold by Slater's Conditions, which are satisfied by our assumption of separability and the form of $\mathcal{P}1$.

Note that we can find the dual function $g(\boldsymbol{\lambda})$ by solving minimizing the lagrangian over $\mathbf{w}, b$ (an unconstrained convex opt. problem). Taking partial derivatives we have that

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}}(\mathbf{w}^*, b, \lambda) = 0 \qquad \Longrightarrow \qquad \mathbf{w}^* = \sum_{i=1}^{m} \lambda_i y_i \mathbf{x}_i \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial b}(\mathbf{w}, b^*, \lambda) = 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{m} \lambda_i y_i = 0 \tag{8}$$

In addition, since strong duality holds, the complementary slackness (C-S) conditions are

$$\lambda_i^* \cdot \left(1 - y_i \left[\mathbf{w}^{*T}\mathbf{x}_i + b^*\right]\right) = 0 \quad i = 1, \ldots, m \tag{9}$$

Substituting the results from (7) and (8) into the Lagrangian, the dual function

$$g(\lambda) = \sum_{i=1}^{l} \lambda_i - \frac{1}{2}\sum_{i,j}^{l} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \tag{10}$$

gives us the dual problem

$$\mathcal{D}1: \quad d^* := \max_{\lambda \geq 0} \sum_{i=1}^{l} \lambda_i - \frac{1}{2}\sum_{i,j}^{l} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \quad s.t. \quad \sum_{i=1}^{l} \lambda_i y_i = 0 \tag{11}$$

$$= \max_{\lambda} \; -\frac{1}{2}\lambda^T \mathbf{Q}\lambda + \mathbf{1}^T\lambda \quad s.t. \quad \lambda^T \mathbf{y} = 0, \quad \lambda \geq \mathbf{0} \tag{12}$$

where$(\mathbf{Q})_{ij} = y_i y_j \mathbf{x}_i^T \mathbf{x}_j$. We can also write $\mathbf{Q} = \mathbf{y}\mathbf{y}^T \circ \mathbf{X}\mathbf{X}^T$, where $\circ$ denotes the hadamard product.

Note that (12) is a quadratic program (QP), and for reasonably sized data it can be solved using a standard QP-solver (e.g. CV, Mosek, SeDuMi, etc.). Methods for very large l or n data often require specific algorithms for SVM. Well-known implementations include LIBSVM and LIBLINEAR.

Once λ^* has been found, the optimal primal variables $(\mathbf{w}^*, b^*)$ are given by

$$\mathbf{w}^* = \sum_{i=1}^{m} \lambda_i^* y_i \mathbf{x}_i \tag{13}$$

Let $\mathcal{I} = \{i : y_i \left[\mathbf{w}^{*T}\mathbf{x}_i + b^*\right] = 1\}$. By the complementary slackness conditions (9), we have that

$$b^* = \frac{1}{y_i} - \mathbf{w}^{*T}\mathbf{x}_i \qquad \text{for } i \in \mathcal{I} \tag{14}$$

$$= \frac{1}{N_{SV}} \sum_{k \in \mathcal{I}}^{N_{SV}} y_k - \mathbf{w}^{*T}\mathbf{x}_k \tag{15}$$

where N_{SV} is the number of support vectors. The average is taken since the calculation of b^* is numerically sensitive. Note that we have derived the OCSH paratmers $(\mathbf{w}^*, b^*)$ as a linear combination of the training data points and that they are calculated using only the Support Vectors (SVs)–this follows by the complementary slackness conditions. Finally, our optimal decision rule, based on the data and the OCSH, becomes

$$f^*(x) = \text{sgn}\left(\mathbf{w}^{*T}\mathbf{x} + b^*\right) = \text{sgn}\left(\sum_{i=1}^{m} \lambda_i^* y_i \mathbf{x}_i^T \mathbf{x} + b^*\right) \tag{16}$$

However, it is naive to think that data can always be separated by a hyperplane, and thus an extension follows for overlapping classes. In section 2, we consider a non-linear boundaries.

1.2 Non-Separable Case: The Linear Soft-Margin Classfier

If the classes overlap, we will not be able to find a feasible $(\mathbf{w}, b)$ pair that satisfies the class constraints in (4). Thus, we introduced the idea of a *soft-margin* which allows data points to lie within the margins.

Idea: Introduce *slack variables* ξ_i $(i = 1, \ldots, l)$ into the constraints and penalize them in objective. The new problem becomes

$$\mathcal{P}2: \quad \min_{\mathbf{w}, b, \boldsymbol{\xi}} \; \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m} \xi_i \tag{17}$$

$$s.t. \quad y_i \left[\mathbf{w}^T\mathbf{x}_i + b\right] \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0 \quad i = 1, \ldots, m$$

I'll note that many generalizations of this form have been made by exponetiating the ξ or replacing $\|w\|_2^2$ with an $\|w\|_1$ and the like. We won't address those cases here. We return to solving $\mathcal{P}2$ using the same duality ideas that we used in solving $\mathcal{P}1$.

First, the Lagrangian for $\mathcal{P}2$ is

$$\text{Ł}(\mathbf{w}, b, \boldsymbol{\xi}, \lambda, \boldsymbol{\gamma}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\lambda_i\left(1 - \xi_i - y_i\left[\mathbf{w}^T\mathbf{x}_i + b\right]\right) - \sum_{i=1}^{m}\gamma_i \cdot \xi_i \tag{18}$$

First, we'll look at the dual problem by performing the inner minimization

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}}(\mathbf{w}^*, b, \boldsymbol{\xi}, \lambda, \boldsymbol{\gamma}) = 0 \qquad \Longrightarrow \qquad \mathbf{w}^* = \sum_{i=1}^{m}\lambda_i y_i \mathbf{x}_i \tag{19}$$

$$\frac{\partial\mathcal{L}}{\partial b}(\mathbf{w}, b^*, \boldsymbol{\xi}, \lambda, \boldsymbol{\gamma}) = 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{m}\lambda_i y_i = 0 \tag{20}$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i^*}(\mathbf{w}, b, \boldsymbol{\xi}^*, \lambda, \boldsymbol{\gamma}) = 0 \qquad \Longrightarrow \qquad C = \lambda_i + \gamma_i, \qquad i = 1, \ldots, m \tag{21}$$

Does Strong Duality hold? Yes. To see this, pick any $\mathbf{w}, b$ and let

$$a = \min_{1\le i\le m} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

Define $\xi_i = 1 - a\ \forall i$ then strict feasibility (which includes active affine constraints) holds and slater's theorem implies strong duality. Again, we will take a look at the C-S conditions

$$\lambda_i^*\left(1 - \xi_i^* - y_i\left[\mathbf{w}^{*T}x_i + b^*\right]\right) = 0, \qquad i = 1, \ldots, m \tag{22}$$

$$\gamma_i^* \cdot \xi_i^* = (C - \lambda_i^*)\xi_i^* = 0, \qquad i = 1, \ldots, m \tag{23}$$

Plugging in (19), (20), and (21) give us back the same dual function as before (10). However, (21) and $\boldsymbol{\gamma}, \lambda \ge \mathbf{0}$ institute a new requirement on the dual problem so that it becomes

$$\begin{aligned} \mathcal{D}2: \quad \max_{\lambda} \ & \sum_{i=1}^{m}\lambda_i - \frac{1}{2}\sum_{i,j}^{l}\lambda_i\lambda_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \qquad (24)\\ s.t. \quad & \sum_{i=1}^{m}\lambda_i y_i = 0 \\ & 0 \le \lambda_i \le C, \qquad i = 1, \ldots, m \\ = \max_{\lambda} \ & -\frac{1}{2}\boldsymbol{\lambda}^T\mathbf{Q}\boldsymbol{\lambda} + \mathbf{1}^T\boldsymbol{\lambda} \qquad (25)\\ s.t. \quad & \boldsymbol{\lambda}^T\mathbf{y} = 0 \\ & \mathbf{0} \le \boldsymbol{\lambda} \le C \cdot \mathbf{1} \end{aligned}$$

still remaining a QP.

Up to this point, we have made no mention of what value C takes and what it represents. The penalty parameter C represents the trade-off between margin size and the number of misclassified points. For example, taking $C = \infty$ requires that all point be correctly classified (this may be infeasible). On the other hand, taking $C = 0$, tailors (17) to focus on maximizing the margin. C is thus usually chosen using cross-validation to minimize estimated generalization error (aka empirical risk).

1.2.1 Support Vectors

After (25) is solved, the slackness conditions (22) and (23) imply three scenarios for the training data points $\mathbf{x}_i$ and the lagrange multipliers λ_i associated with their classification constraints:

1. ($\lambda_i = 0$ and $\xi_i = 0$): Then, the data point x_i has been correctly classified.

2. ($0 < \lambda_i < C$): By (21) and (23), $\xi_i \equiv 0$, which implies that $y_i \left[\mathbf{w}^T \mathbf{x}_i + b\right] = 1$. Thus, $\mathbf{x}_i$ is a support vector. Note that the support vectors that satisfy $0 \leq \lambda_i < C$ are the *unbounded* or *free* support vectors.

3. ($\lambda_i = C$): Then by (22), $y_i \left[\mathbf{w}^T \mathbf{x}_i + b\right] = 1 - \xi_i$, $\xi_i \geq 0$, and $\mathbf{x}_i$ is a SV. Note that the SVs with $\lambda_i = C$ are *bounded* support vectors; that is, they lie inside the margin. Furthermore, for $0 \leq \xi_i < 1$, $\mathbf{x}_i$ is correctly classified, but if $\xi_i \geq 1$, $\mathbf{x}_i$ is misclassified.

2 The Nonlinear Classifier

2.1 Feature Space

Although the soft-margin linear classifier allowed us to consider linearly non-separable classes, its scope is still limited since data might (and tend to) have nonlinear hypersurfaces that separate them. The idea here is to map our data $\mathbf{x} \in \mathbb{R}^n$ into some high-dimensional feature space F via a mapping $\mathbf{\Phi}$ ($\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^f$), and then solve the linear classification problem in this space.

How does this affect (24) ? We just replace x_i's with $\mathbf{\Phi}_i = \mathbf{\Phi}(\mathbf{x}_i)$, but actually there's a better way to solve the problem without ever actually mapping the points explicitly. We will use the *kernel trick* to greatly reduce the number of necessary computations.

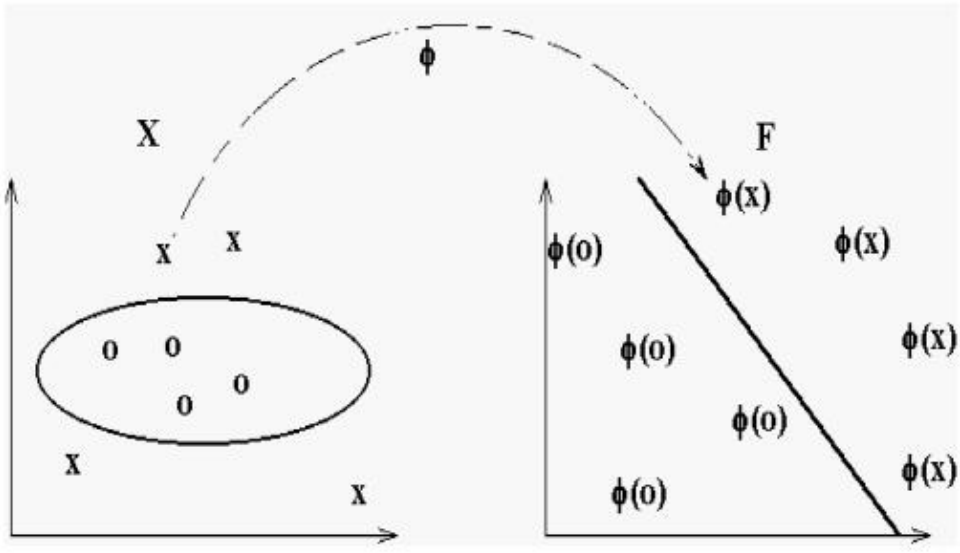

Figure 2: Nonlinear Classifier (Picture also from Wikipedia)

2.2 Nonlinear Hypersurfaces

If we want to create a hyperplane, in the high-dimensional feature space F we will need to be able to carry out inner-products. However, depending on the dimension of F, this could become very costly, first in mapping the data and secondly in taking inner products. Thus, we introduce the concept of the kernel function. A function which allows us to take the inner product of transformed (by $\mathbf{\Phi}$) data without actually transforming the data. Huh? A *kernel function* satisfies the following

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{\Phi}^T(\mathbf{x}_i)\mathbf{\Phi}(x_j) \tag{26}$$

The most popular kernels are

Function ($K(\mathbf{x}, \mathbf{x}_i)$)	Type of Classifier	Positive-Definite
$\mathbf{x}^T\mathbf{x}_i$	Linear, dot product	Cond. PD
$\left(\mathbf{x}^T\mathbf{x}_i + b\right)^d$	Polynomial Degree d	PD
$\exp\{\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}_i)\}$	Gaussian Radial Basis Function†	PD
$\tanh(\mathbf{x}^T\mathbf{x}_i + b)^*$	Multilayer Perceptron	Cond. PD
$\left(\sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2 + \beta}\right)^{-1}$	Inverse Multiquadratic Function	PD

* may not be a valid kernel for some b.
† $\mathbf{\Sigma}$ is usually taken to be isotropic. I.e. $\mathbf{\Sigma} = \sigma^2\mathbf{I}$

2.3 The Kernel Trick

After a feature space, and accordingly its kernel, has been selected, the linear classification problem *with soft-margins* becomes

$$\begin{aligned}
&\max_{\lambda} \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j}^{l} \lambda_i\lambda_j y_i y_j \mathbf{\Phi}_i^T\mathbf{\Phi}_j && (27)\\
&s.t. \quad \sum_{i=1}^{m} \lambda_i y_i = 0\\
&\qquad 0 \le \lambda_i \le C, \qquad i = 1, \ldots, m\\
&= \max_{\lambda} \sum_{i=1}^{m} \lambda_i - \frac{1}{2}\sum_{i,j}^{l} \lambda_i\lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) && (28)\\
&s.t. \quad \sum_{i=1}^{m} \lambda_i y_i = 0\\
&\qquad 0 \le \lambda_i \le C, \qquad i = 1, \ldots, m\\
&= \max_{\lambda} -\frac{1}{2}\lambda^T\mathbf{H}\lambda + \mathbf{1}^T\lambda && (29)\\
&s.t. \quad \lambda^T y = 0\\
&\qquad \mathbf{0} \le \lambda \le C \cdot \mathbf{1}
\end{aligned}$$

where $H_{ij} = y_i y_j K(x_i, x_j)$. Herein lies the importance of the Gram matrix $\mathbf{G}$, where $(\mathbf{G})_{ij} = K(x_i, x_j)$, because if G is positive definite (PD), then the matrix $[y_i y_j \mathbf{G}_{ij}]$ will be PD and the optimization problem above will be convex and have a unique solution.

3 Additional Exercises

Question 1: (SVM w/o Bias Term) Formulate a soft-margin (i.e. with ξ_i) SVM without the bias term, i.e. $f(x) = w^T x$. Derive the saddle point conditions, KKT conditions and the dual.

Question 2: (SVM - Primal, Dual QP's) Both the primal and dual forms of the soft-margin SVM are QP's.

1. In the linear case, what problems parameters determine the computational complexity of the primal and dual problems? In what regimes, would one be faster than the other?
2. Consider now the case when we use non-linear feature mapping. What can be said now?

References

[1] Lipo Wang
Support Vector Machines: Theory and Applications.
Springer, Netherlands, 2005