CS5339 Lecture Notes #2: Support Vector Machine

Jonathan Scarlett

March 30, 2021

Useful references:

- Blog post by Jeremy Kun[1]
- MIT lecture notes,[2] lecture 3
- Chapter 7 of Bishop's "Pattern Recognition and Machine Learning" book
- Chapter 15 of "Understanding Machine Learning" book
- Wikipedia page on Support Vector Machine
- Supplementary notes lec3a.pdf

1 Binary Classification

Recap of the classification problem:

- The *data set* is given by $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_{t=1}^n$ where $\mathbf{x}_t \in \mathbb{R}^d$ are the input vectors and $y_t \in \{-1, +1\}$ are the targets/labels
- A *classifier* is a function $f : \mathbb{R}^d \to \{-1, +1\}$ that takes $\mathbf{x}$ as input and tries to predict the corresponding label y.
- *Linear classifiers* are those in the set

$$\mathcal{F} = \big\{ f \,:\, f(\mathbf{x}) = \mathrm{sign}(\mathbf{x}^T \boldsymbol{\theta}) \text{ for some } \boldsymbol{\theta} \in \mathbb{R}^d \big\}.$$

- The data set $\mathcal{D}$ is said to be *linearly separable* if there exists a linear classifier (i.e., a choice of $\boldsymbol{\theta}$) that classifies everything in the data set $\mathcal{D}$ correctly. We will continue with this assumption initially, but will shortly drop it.

Margin of a classifier.

[1]`http://jeremykun.com/2017/06/05/formulating-the-support-vector-machine-optimization-problem/`

[2]`http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-867-machine-learning-fall-2006/lecture-notes/`

- Recall that we defined the margin corresponding to $\boldsymbol{\theta}$ as $\gamma_{\text{geom}} = \frac{\gamma}{\|\boldsymbol{\theta}\|}$, where

$$\gamma = \min_{t=1,\dots,n} y_t \boldsymbol{\theta}^T \mathbf{x}_t.$$

- At least intuitively, a larger margin should lead to a "more robust" classifier.

2 Maximum Margin Classifier – Initial Formulation

Maximizing the margin.

- We can write down the maximum margin classifier as an optimization problem:

$$\text{maximize}_{\boldsymbol{\theta},\gamma} \ \frac{\gamma}{\|\boldsymbol{\theta}\|} \ \text{ subject to } \ y_t \boldsymbol{\theta}^T \mathbf{x}_t \ge \gamma, \ \ \forall t = 1, \dots, n.$$

- For convenience, we rewrite the maximization as minimizing the inverse:

$$\text{minimize}_{\boldsymbol{\theta},\gamma} \ \frac{\|\boldsymbol{\theta}\|}{\gamma} \ \text{ subject to } \ \frac{y_t \boldsymbol{\theta}^T \mathbf{x}_t}{\gamma} \ge 1, \ \ \forall t = 1, \dots, n.$$

  We have also divided both sides by $\gamma > 0$ in each constraint.

- Then, since everything depends on $\boldsymbol{\theta}$ and γ only through $\frac{\boldsymbol{\theta}}{\gamma}$, we can just define $\widetilde{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta}}{\gamma}$ and form the equivalent problem

$$\text{minimize}_{\widetilde{\boldsymbol{\theta}}} \ \|\widetilde{\boldsymbol{\theta}}\| \ \text{ subject to } \ y_t \widetilde{\boldsymbol{\theta}}^T \mathbf{x}_t \ge 1, \ \ \forall t = 1, \dots, n.$$

- Finally, maximizing a quantity is equivalent to maximizing its square, so we write yet another equivalent form (let's also drop the tilde on $\widetilde{\boldsymbol{\theta}}$ for simpler notation):

$$\text{minimize}_{\boldsymbol{\theta}} \ \frac{1}{2}\|\boldsymbol{\theta}\|^2 \ \text{ subject to } \ y_t \boldsymbol{\theta}^T \mathbf{x}_t \ge 1, \ \ \forall t = 1, \dots, n. \tag{1}$$

  The solution $\boldsymbol{\theta}$ to this problem is a basic version (i.e., one only suited to linearly separable data) of the *support vector machine* (SVM) classifier.

Uniqueness of the solution:

- Claim. The solution to the optimization problem (1) is unique.
- Proof:
  - Suppose, to the contrary, there were two solutions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Such solutions clearly need to satisfy $\|\boldsymbol{\theta}_1\| = \|\boldsymbol{\theta}_2\|$; let's give this norm a name V^*.
  - Now consider the alternative choice $\bar{\boldsymbol{\theta}} = \frac{1}{2}\boldsymbol{\theta}_1 + \frac{1}{2}\boldsymbol{\theta}_2$. The triangle inequality gives

$$\|\bar{\boldsymbol{\theta}}\| \le \frac{1}{2}\big(\|\boldsymbol{\theta}_1\| + \|\boldsymbol{\theta}_2\|\big) = V^*, \tag{2}$$

    so the norm cannot be any larger than V^*. Also, since the constraints are linear and satisfied by both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, they are satisfied by $\bar{\boldsymbol{\theta}}$. Since V^* is the smallest possible norm by definition, we conclude that (2) can only hold with equality: $\|\bar{\boldsymbol{\theta}}\| = V^*$. Substituting $\bar{\boldsymbol{\theta}} = \frac{1}{2}\boldsymbol{\theta}_1 + \frac{1}{2}\boldsymbol{\theta}_2$ and squaring gives $\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|^2 = 4(V^*)^2$.

  – Next, using expansion of the square, we have

$$\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|^2 = \|\boldsymbol{\theta}_1\|^2 + 2\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\rangle + \|\boldsymbol{\theta}_2\|^2 = 2((V^*)^2 + \langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\rangle)$$
$$\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 = \|\boldsymbol{\theta}_1\|^2 - 2\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\rangle + \|\boldsymbol{\theta}_2\|^2 = 2((V^*)^2 - \langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\rangle)$$

and adding these equations together gives

$$\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|^2 + \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 = (4V^*)^2.$$

But we already showed $\|\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2\|^2 = 4(V^*)^2$, so these can only be consistent if $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|^2 = 0$, meaning $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

3 Support Vector Machine – Towards a General Formulation

Adding an offset parameter

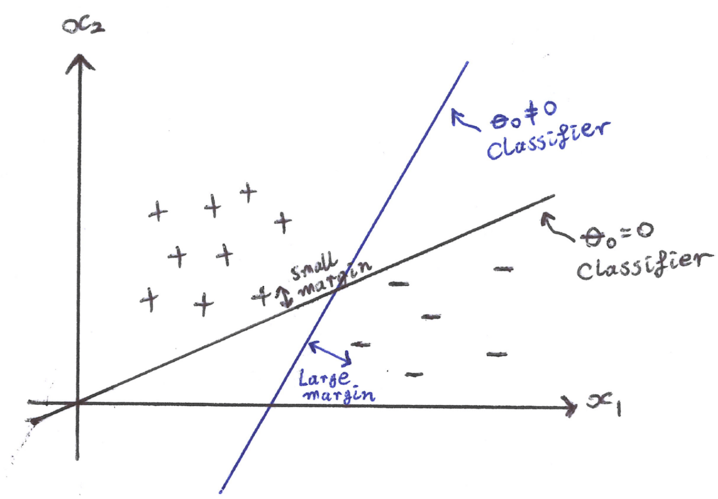
- Let's slightly generalize linear classifiers as follows:

$$\mathcal{F} = \big\{ f \,:\, f(\mathbf{x}) = \mathrm{sign}(\boldsymbol{\theta}^T\mathbf{x} + \theta_0) \text{ for some } \boldsymbol{\theta} \in \mathbb{R}^d, \theta_0 \in \mathbb{R} \big\}.$$

  The previous formulation corresponds to choosing $\theta_0 = 0$. This extra parameter is called the *offset* or *bias* of the classifier.

  – We will usually refer to these as *linear classifiers* as well, though the more precise terminology would be *affine classifiers*.

- The added flexibility of the offset parameter can improve the margin:



- The inclusion of θ_0 changes the SVM formulation slightly:

$$\text{minimize}_{\boldsymbol{\theta},\theta_0} \ \frac{1}{2}\|\boldsymbol{\theta}\|^2 \ \text{ subject to } \ y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \geq 1, \ \ \forall t = 1, \dots, n. \tag{3}$$

- Notes:
  - θ_0 only appears in the constraints, not the objective
  - If we were to apply (1) to the modified domain $\tilde{\mathbf{x}}_t = [\mathbf{x}_t^T \ 1]^T$ with $[\boldsymbol{\theta}^T \ \theta_0]^T$ in place of $\boldsymbol{\theta}$, the parameter θ_0 would affect both the constraints and objective. The two formulations are **not** equivalent; it is only (3) that is correct for maintaining the maximum-margin interpretation.

Allowing mis-classified examples.

- Most data sets are not linearly separable (even with the flexibility of the offset θ_0).
- Intuition on the general SVM: Allow margin violations and mis-classified examples, but pay a penalty for them.
  - Since violations are allowed, we refer to this as the *soft-margin SVM*. The previous formulation with no violations is called the *hard-margin SVM*.
- The optimization formulation:

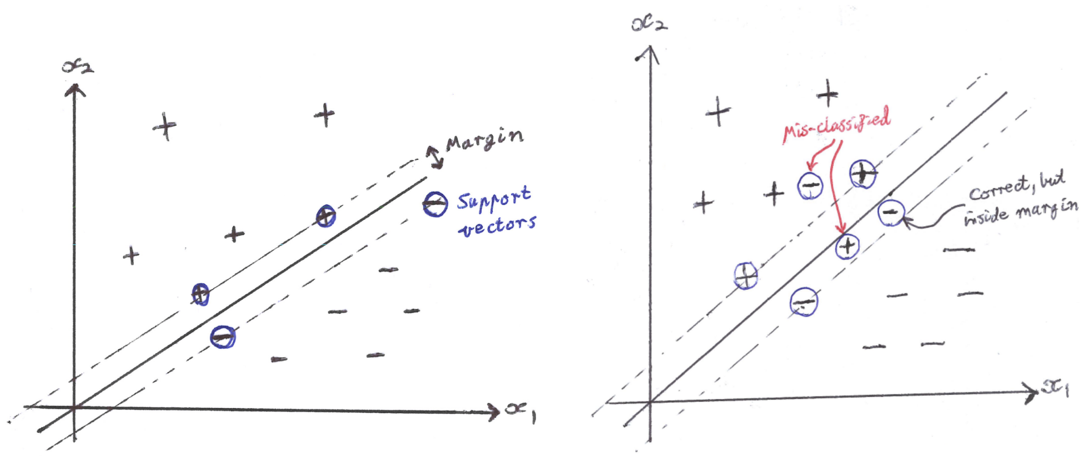
$$\text{minimize}_{\boldsymbol{\theta},\theta_0,\boldsymbol{\zeta}} \ \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{n}\zeta_t \ \text{ subject to } \ y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \geq 1 - \zeta_t \ \text{ and } \ \zeta_t \geq 0, \quad \forall t \tag{4}$$

  where $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ is an extra set of optimization variables called *slack variables*, and C is a parameter controlling how the two terms in the objective are weighted.

- Remarks.
  1. If $\zeta_t = 0$, we still satisfy $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \geq 1$ as before. If $\zeta_t > 0$, we are no longer "within the margin". If $\zeta_t > 1$, we don't even classify $\mathbf{x}_t$ correctly (see below).
  2. As C grows very large, the optimal slack variables ζ_t will become closer to zero (why?), and we simply recover the maximum margin rule (if the data set is linearly separable). But if C gets small, more and more margin violations are permitted.
  3. Overall, C controls the trade-off between having a large margin ($\frac{1}{2}\|\boldsymbol{\theta}\|^2$ term) and few margin violations ($\sum_{t=1}^{n}\zeta_t$ term).
- In practice, C might require some tuning (e.g., via cross-validation, to be covered later).

So what is a support vector?

- The *support vectors* are the samples $(\mathbf{x}_t, y_t)$ falling into any of the following categories:
  - Those that lie exactly on the margin
  - Those that violate the margin constraint, but not enough to be mis-classified
  - Those that are mis-classified



- An example (separable case on left, non-separable on right):
- If we apply the SVM to a reduced data set consisting of *only* the support vectors, we get back the *exact same classifier*.
  - We will skip a formal proof of this fact here; it can be shown using techniques that we introduce for a "dual" SVM formulation later in the course.
  - The intuition (separable case): Attaining the maximum margin can be viewed as stretching out a "slab" (parallel to the decision boundary) until some data points are "hit". Even if we remove those that were not hit, we still hit the same ones that were kept.

Yet another equivalent formulation.

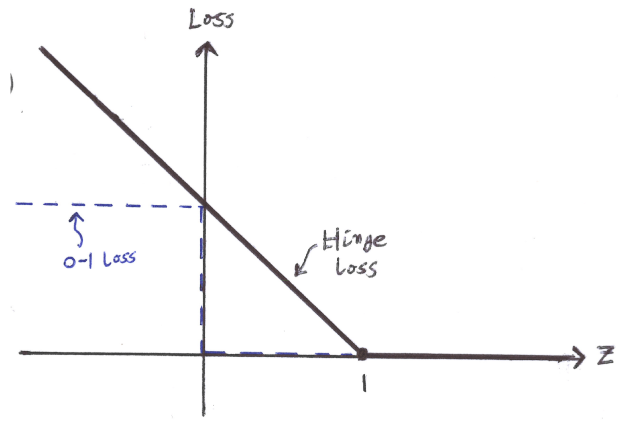
- Claim. The optimization (4) is equivalent to the *unconstrained* problem

$$\text{minimize}_{\boldsymbol{\theta},\theta_0} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{t=1}^{n}\big[1 - y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0)\big]_+, \tag{5}$$

  where $[z]_+ = \max\{0, z\}$.

  - Proof. (i) If $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) > 1$, then we have $[1 - y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0)]_+ = 0$ and pay no penalty, just like in (4). (ii) If $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \le 1$, then we have $[1 - y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0)]_+ = 1 - y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0)$, which matches the penalty ζ_t in (4).
  - (To properly establish the last part of this argument, try to convince yourself that whenever $\zeta_t > 0$ the constraint $y_t(\boldsymbol{\theta}^T\mathbf{x}_t + \theta_0) \ge 1 - \zeta_t$ holds with equality.)

- The function $\mathrm{Loss_h}(z) = [1 - z]_+$ is referred to as the *hinge loss*:



- So (5) can be interpreted as balancing the *total hinge loss* with the *regularization term* $\frac{1}{2}\|\boldsymbol{\theta}\|^2$. The terminology "regularization" will be discussed more in later lectures.

- <u>A note on computation.</u>
  - The above SVM formulations are so-called *convex optimization* problems (to be defined formally in a later lecture), for which there exist general-purpose solvers that can efficiently find the solution numerically. For instance, (1) minimizes a quadratic function subject to linear constraints.
  - By contrast, if we tried replacing the hinge loss by the 0-1 loss, we would have an optimization formulation that is extremely hard to solve in general (specifically, NP-hard).

- **In a later lecture:**
  - A completely different yet equivalent optimization formulation called the *dual expression* (the ones we have presented so far are called *primal expressions*).
  - A way to produce non-linear classifiers via the "kernel trick".