CMSE 820 HW9

Hao Lin

November 17, 2018

1 Question 1

1.1 Hard margin SVM on a separable dataset

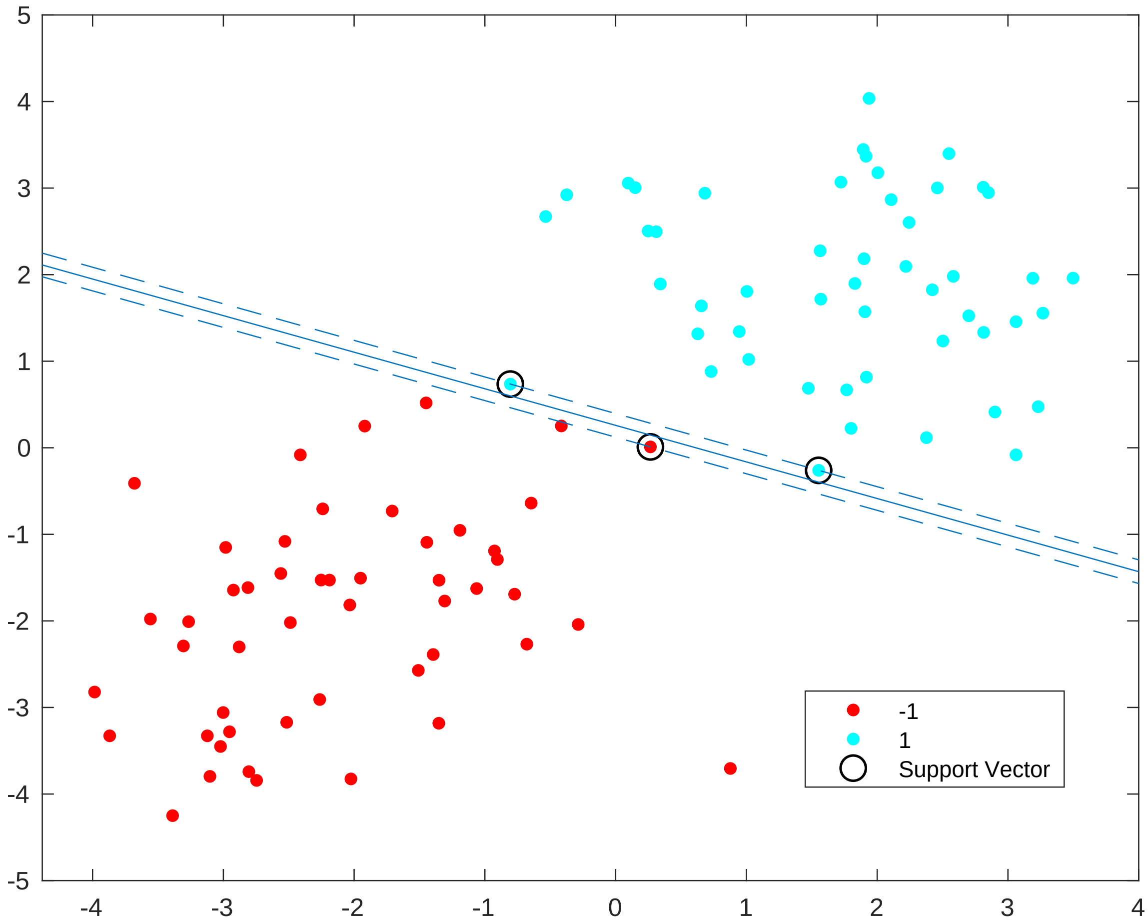

Figure 1: Hard margin SVM on a separable dataset. The colors indicate the *original* label in the dataset.

1.2 Hard margin SVM on non-separable dataset

After making the change in the dataset, the dataset becomes non-separable, so hard margin SVM yields no solution.

1.3 Soft margin SVM on a non-separable dataset

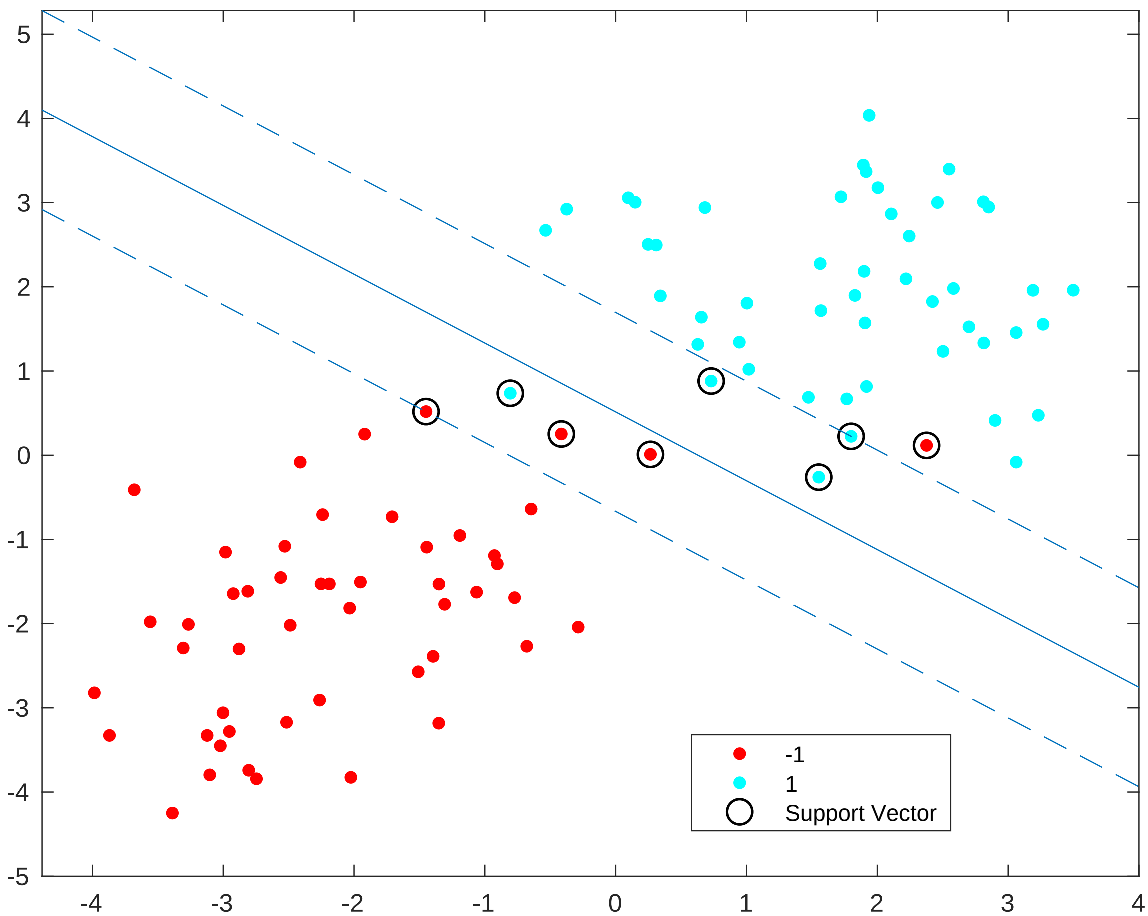

Figure 2: Soft margin SVM on a non-separable dataset. $\gamma = 1$. The colors indicate the *original* label in the dataset.

Note that evening using soft margin SVM with $\gamma = 1$, the outlier red dot (-1) still lies outside the margin and is misclassified as having label 1.

2 Question 2

2.1 Hard margin SVM

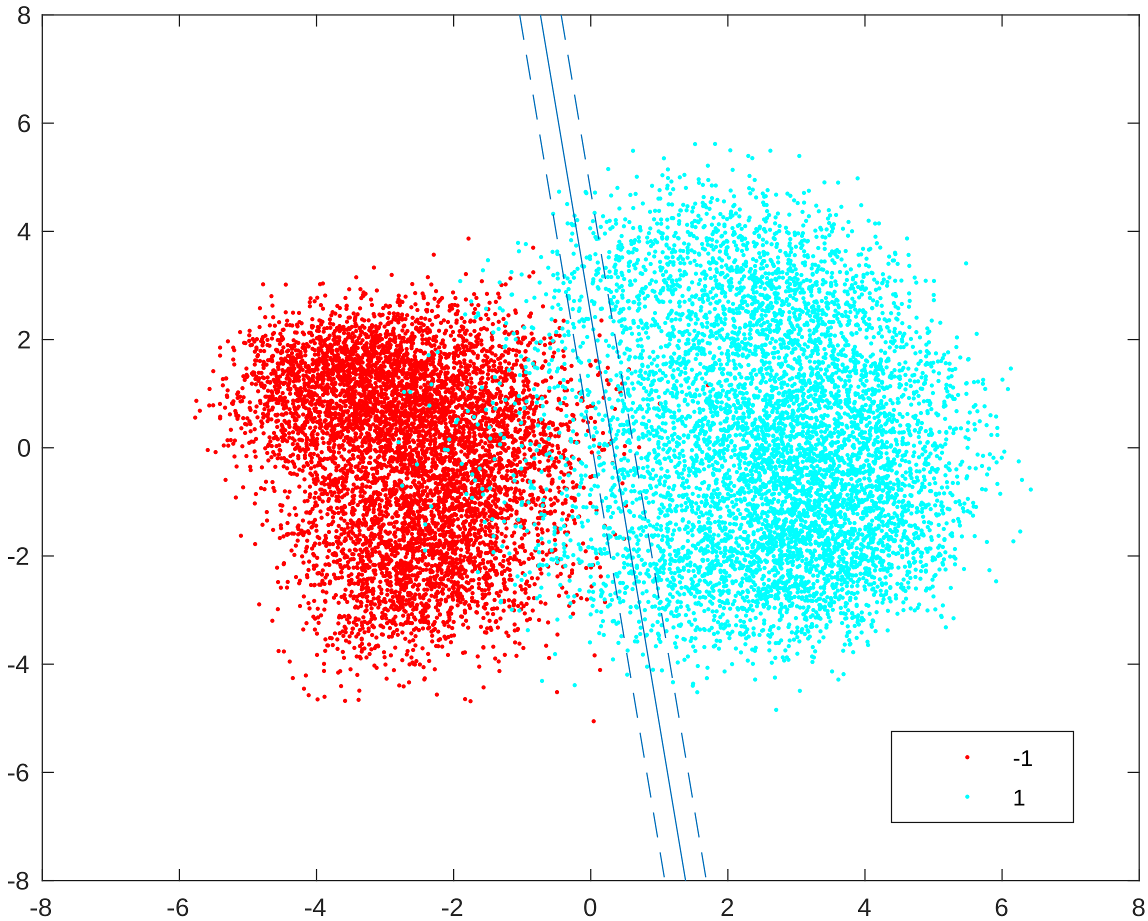

Figure 3: Hard margin SVM on the *original* dataset. The solid line is the *intersection* of the SVM hyperplane and the plane spanned by the top 2 PCs. The colors indicate the *original* label in the dataset. The two axes are the top 2 PCs.

I ran Hard margin SVM on the *original* dataset. The original data was then projected onto the top 2 PCs. **The solid line represents the intersection of the SVM hyperplane and the plane spanned by the top 2 PCs**. Although the data are not separable along the top 2 PCs, but they are indeed **separable in the full 400-dimensional space**. The accuracy given by the hard margin SVM model is thus **100%**.

2.2 Soft margin SVM

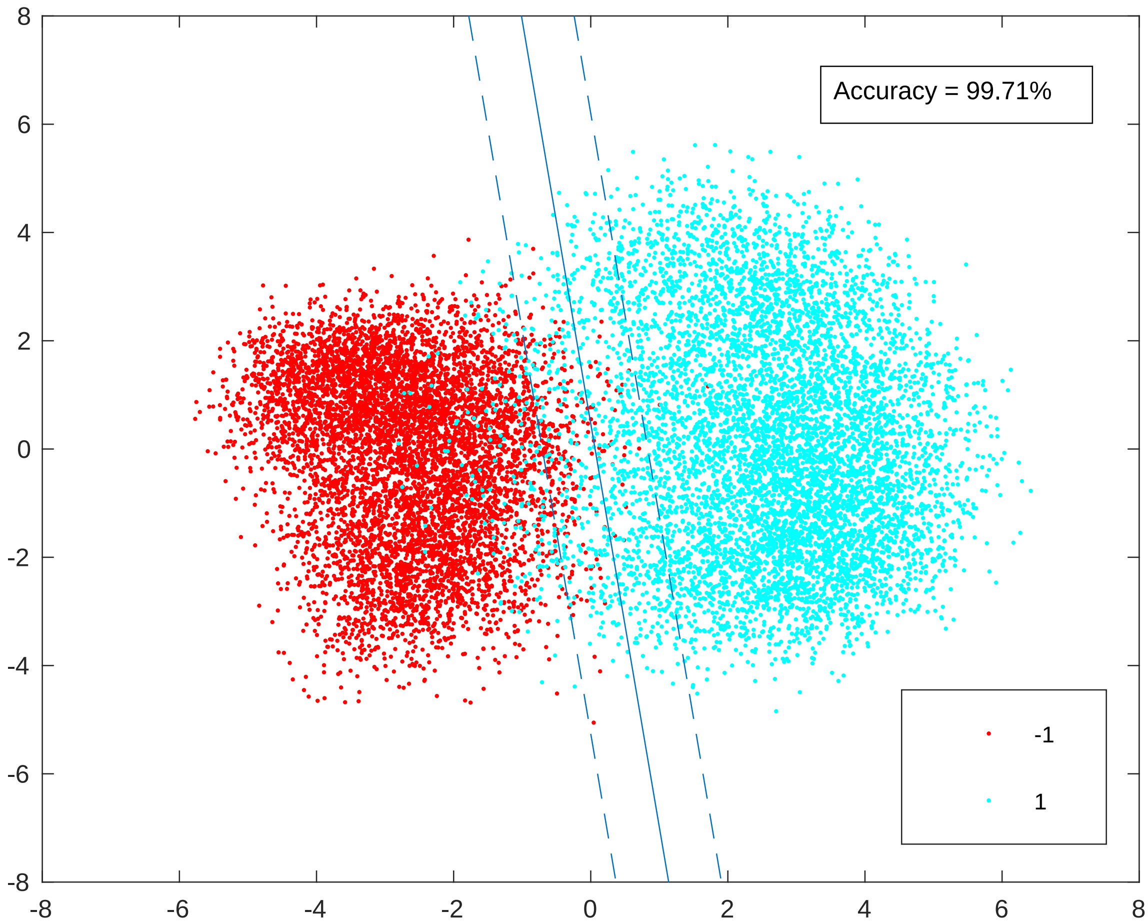

Figure 4: Soft margin SVM ($\gamma = 0.04$) on the *original* dataset. The solid line is the *intersection* of the SVM hyperplane and the plane spanned by the top 2 PCs. The colors indicate the *original* label in the dataset. The two axes are the top 2 PCs.

I ran soft margin SVM on the *original* dataset and used **2-fold cross-validation** to find the optimal γ. The optimal γ is 0.04 and the accuracy is **99.71%**. The original data was then projected onto the top 2 PCs. **The solid line represents the intersection of the SVM hyperplane and the plane spanned by the top 2 PCs**.

2.3 SVM with Gaussian kernel

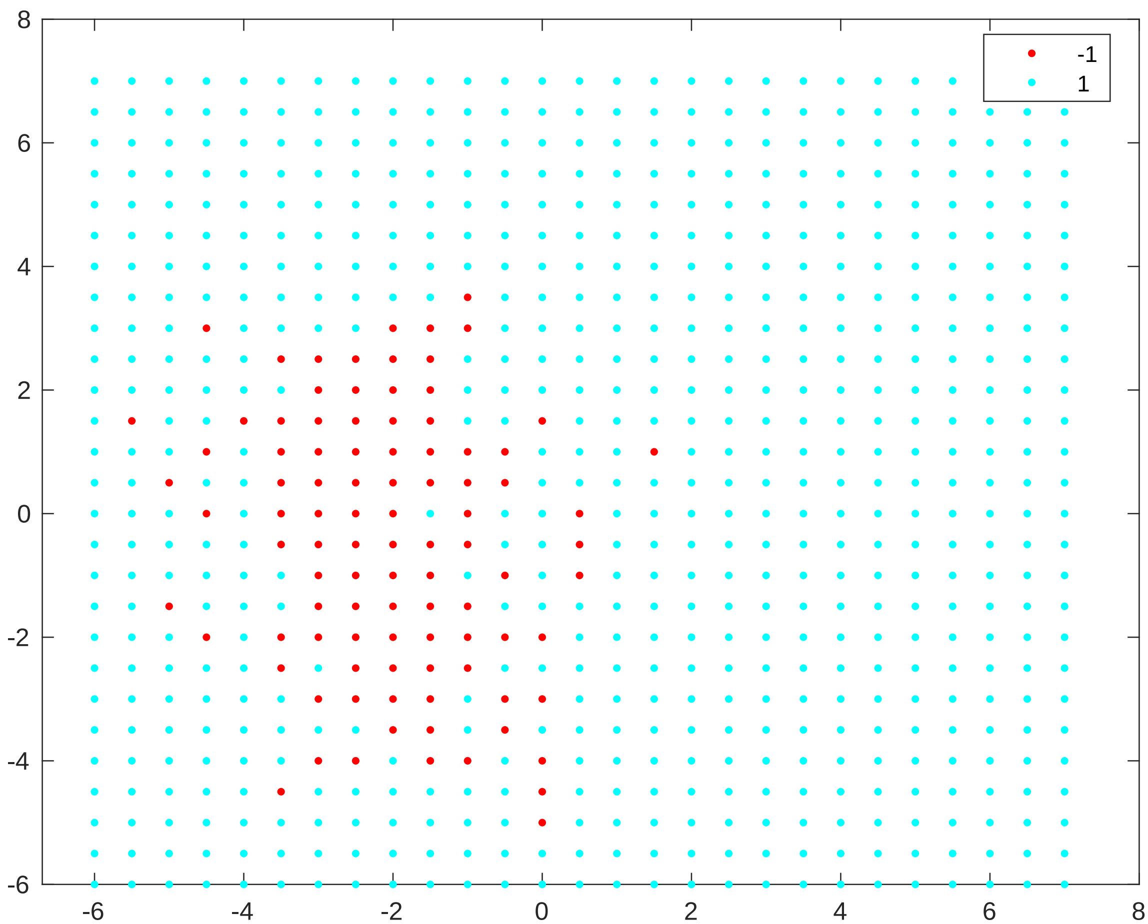

Figure 5: Predictions of the SVM model with a Gaussian kernel ($\sigma = 0.3$) on a grid spanned along the top 2 PCs.

I ran SVM with a Gaussian kernel on the top 2 PCs (rather than the whole original dataset) and cross-validated it to find the optimal $\sigma = 0.3$. The accuracy is merely **88.51%**, much lower than the previous two SVM models. This is because, in this case, I performed PCA before running SVM and a lot of the information is lost.

3 Question 3

Soft Margin SVM
(1) Regularized risk form:

$$\min_{w,b} \sum_i [1 - y_i(\langle w, x_i \rangle + b)]_+ + \frac{\lambda}{2}\|w\|^2.$$

(2) Primal form:

$$\begin{aligned}
&\min_{w,b} \gamma \sum_i \xi_i + \frac{1}{2}\|w\|^2 \\
&\text{subject to}\, \xi_i \geq 0, \\
&y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i.
\end{aligned}$$

Show that they are equivalent with $\lambda = 1/\gamma$.

Proof: Substituing $\lambda = 1/\gamma$, the regularized risk form problem can be rewritten as

$$\min_{w,b} \gamma \sum_i \max\{0, 1 - y_i(\langle w, x_i \rangle + b)\} + \frac{1}{2}\|w\|^2.$$

Let $\{w_1, b_1\}$ and $\{w_2, b_2\}$ denote respectively the optimal solutions to the regularized risk form problem and the primal form problem respectively.
The Lagrangian for the primal problem reads

$$L(w, b, \xi_i, \alpha_i, \lambda_i) = \gamma \sum_i \xi_i + \frac{1}{2}\|w\|^2 + \sum_i \alpha_i[1 - \xi_i - y_i(\langle w, x_i \rangle + b)] - \sum_i \lambda_i \xi_i.$$

The complementary slackness yields

$$\forall\, i,\ \lambda_i \xi_i = 0 \text{ and } \alpha_i[1 - \xi_i - y_i(\langle w_2, x_i \rangle + b_2)] = 0,$$

which implies that

$$\forall\, i,\ \xi_i = \max\{0, 1 - y_i(\langle w_2, x_i \rangle + b_2)\}.$$

So the objective function of the primal problem at the optimal solution $\{w_2, b_2\}$ is given by

$$\gamma \sum_i \max\{0, 1 - y_i(\langle w_2, x_i \rangle + b_2)\} + \frac{1}{2}\|w_2\|^2 \geq \gamma \sum_i \max\{0, 1 - y_i(\langle w_1, x_i \rangle + b_1)\} + \frac{1}{2}\|w_1\|^2.$$

The inequality follows from that $\{w_1, b_1\}$ is the optimal solution to the regularized form problem.

On the other hand, given $\{w_1, b_1\}$, we can define $\xi_i' := \max\{0, 1 - y_i(\langle w_1, x_i\rangle + b_1)\}$ such that ξ_i' is feasible for the primal form problem. The optimality of the primal form solution $\{w_2, b_2\}$ gives

$$\gamma \sum_i \max\{0, 1 - y_i(\langle w_2, x_i\rangle + b_2)\} + \frac{1}{2}\|w_2\|^2 \leq \gamma \sum_i \max\{0, 1 - y_i(\langle w_1, x_i\rangle + b_1)\} + \frac{1}{2}\|w_1\|^2.$$

We obtain an equality from the previous two inequalities

$$\gamma \sum_i \max\{0, 1 - y_i(\langle w_2, x_i\rangle + b_2)\} + \frac{1}{2}\|w_2\|^2 = \gamma \sum_i \max\{0, 1 - y_i(\langle w_1, x_i\rangle + b_1)\} + \frac{1}{2}\|w_1\|^2.$$

Moreover, since the soft margin SVM problem is a strictly convex optimization problem, the solution is unique, i.e.,

$$w_1 = w_2 \text{ and } b_1 = b_2.$$

Thus, the two forms of problems are equivalent.

4 Question 4

Constrained form:

$$\min_x f(x) \text{ subject to } h(x) \leq t.$$

Lagrange form:

$$\min_x f(x) + \lambda h(x).$$

Equivalent?

Solution: First of all, they are not equivalent in general. A counterexample: If we let $f(x) = x$, $h(x) = -x$ and $t = 1$, then there exists no λ that can help the Lagrange form to yield the same solution.

4.1 Given a t, what should λ be such that the Lagrange form yields the same solution as the constrained form?

We have to impose a few extra conditions: (1) both $f(x)$ and $h(x)$ are at least twice differentiable; (2) the function $g(x) := -f'(x)/h'(x)$ is well-defined and strictly monotonic; (3) the function $j(x) = f''(x) + g(x)h''(x)$ is well-defined and strictly positive.

For any given t, let's denote the optimal solution to the constrained problem as $\tilde{x}(t)$. We claim that when

$$\lambda(t) = -f'[\tilde{x}(t)]/h'[\tilde{x}(t)],$$

the Lagrange form problem is equivalent to the constrained form problem. Taking the derivative of the Lagrangian $L(x) = f(x) - \frac{f'[\tilde{x}(t)]}{h'[\tilde{x}(t)]}h(x)$ and setting it to 0 gives,

$$f'(x) - \frac{f'[\tilde{x}(t)]}{h'[\tilde{x}(t)]}h'(x) = 0.$$

Since the function $-f'(x)/h'(x)$ is strictly monotonic, the only possible solution is given by $x = \tilde{x}(t)$. Indeed, $x = \tilde{x}(t)$ is the optimal solution to the Lagrange problem because $L''(\tilde{x}(t)) > 0$ by assumption (3).

4.2 Given a $\lambda > 0$, what should t be such that the constrained form yields the same solution as the Lagrange form?

Let's denote $x^\star(\lambda)$ as the optimal solution to the Lagrange problem and $\tilde{x}(t)$ as the optimal solution to the constrained problem. We claim that

$$t(\lambda) = h[x^\star(\lambda)].$$

The optimality of $x^\star(\lambda)$ for the Lagrange problem implies that

$$\forall\, x \in \{x : h(x) \le h[x^\star(\lambda)]\},\ f[x^\star(\lambda)] + \lambda h[x^\star(\lambda)] \le f(x) + \lambda h(x) \le f(x) + \lambda h[x^\star(\lambda)],$$

which implies that $f[x^\star(\lambda)] \le f(x)$ for all feasible x in the constrained problem. So $x^\star(\lambda)$ is a solution to the constrained problem as well.