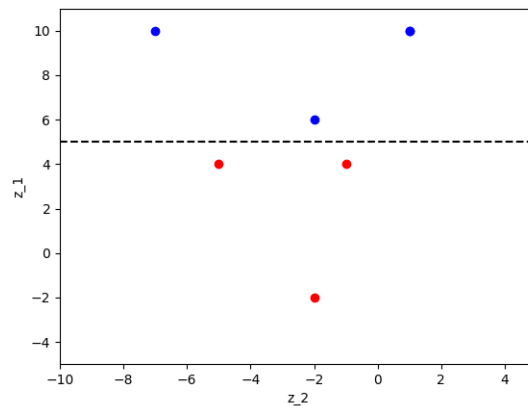


Machine Learning Spring 2019 HW1

Yueh Cheng Liu
National Taiwan University

March 25, 2019

1



The equation of the hyperplane is $z_1 = 5$. Which has the largest margin 2. If the hyperplane slightly rotate clockwise or counter-clockwise, the margin will be shrinking due to the two red points $(-4.5, 4)$ and $(-1, 4)$.

2

$\alpha = (0, 0.704, 0.704, 0.889, 0.259, 0.259, 0)$. x_i with $i = 2, \dots, 6$ are the support vectors.

3

$$\begin{aligned}\sum_{n \in \mathbf{SV}} \alpha_n y_n K(x_n, \mathbf{x}) + b &= \sum_{n \in \mathbf{SV}} \alpha_n y_n (1 + x_n^T \mathbf{x})^2 + b \\ &= \sum_{n \in \mathbf{SV}} \alpha_n y_n (1 + x_n^T \mathbf{x} + (x_n^T \mathbf{x})(x_n^T \mathbf{x})) + b\end{aligned}$$

Let $\mathbf{x} = (x_1, x_2)$ and replace the α_n , y_n , x_n with number.

$$\begin{aligned}\sum_{n \in \mathbf{SV}} \alpha_n y_n (1 + x_n^T \mathbf{x} + (x_n^T \mathbf{x})(x_n^T \mathbf{x})) &= -0.704 \cdot (1 + x_2 + x_2^2) \\ &\quad - 0.704 \cdot (1 + -x_2 + x_2^2) \\ &\quad + 0.889 \cdot (1 + -x_1 + x_1^2) \\ &\quad + 0.259 \cdot (1 + 2x_2 + 4x_2^2) \\ &\quad + 0.259 \cdot (1 + -2x_2 + 4x_2^2) \\ &= -1 - 0.889x_1 + 0.889x_1^2 + 0.664x_2^2\end{aligned}$$

and take a support vector $x_2 = (0, 1)$ to calculate b .

$$\begin{aligned}b &= y_s - \sum_{n=1}^N \alpha_n y_n K(x_n, x_s) \\ &= y_s - \sum_{n=1}^N \alpha_n y_n (1 + x_n^T x_s + (x_n^T x_s)(x_n^T x_s)) \\ &= -1 - \sum_{n=1}^N \alpha_n y_n (1 + x_{n,2} + x_{n,2}^2) \\ &= 0.663\end{aligned}$$

So the curve is $0.889x_1 + 0.889x_1^2 + 0.664x_2^2 - 0.337 = 0$.

4

The curve found in (1) is $2x_2^2 - 4x_1 - 3 = 0$. Which is not the same with (2). The transformation (2) represented by the kernel function is different from the transformation in (1), so the optimal separating hyperplanes for them are different.

5

$$\begin{aligned} ||\tilde{\phi}(x)||^2 &= 1 + \frac{2}{1!}x^2 + \frac{2^2}{2!}x^4 + \dots \\ &= \sum_{n=0}^{\infty} \frac{2^n}{n!}x^{2n} \end{aligned}$$

Let $y = x^2$ and substitute x with y .

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{2^n}{n!}x^{2n} &= \sum_{n=0}^{\infty} \frac{2^n}{n!}y^n \\ &= e^{2y} \\ &= e^{2x^2} \end{aligned}$$

So that

$$e^{x^2} = ||\tilde{\phi}(x)||$$

6

$$\cos(x', x) = \frac{x'^T x}{|x'| |x|}$$

The transform function $\phi(x)$ is $\frac{x}{|x|}$.

7

$$L(R, c, \lambda) = R^2 + \sum_{n=1}^N \lambda_n (||z_n - c||^2 - R^2)$$

8

- Primal feasible: $||z_n - c||^2 \leq R^2$
- Dual feasible: $\lambda_n \geq 0$ for $n = 1, 2, \dots, N$

- Dual-inner optimal:

$$\begin{aligned}\frac{\delta L(R, c, \lambda)}{\delta c} &= 0 = \sum_{n=1}^N \lambda_n (-2z_n + 2c) \\ c \sum_{n=1}^N \lambda_n &= \sum_{n=1}^N \lambda_n z_n\end{aligned}\tag{1}$$

$$\begin{aligned}\frac{\delta L(R, c, \lambda)}{\delta R} &= 0 = 2R - \sum_{n=1}^N \lambda_n 2R \\ R(1 - \sum_{n=1}^N \lambda_n) &= 0\end{aligned}\tag{2}$$

By equation 3, if $\sum_{n=1}^N \lambda_n \neq 0$, then $c = \frac{\sum_{n=1}^N \lambda_n z_n}{\sum_{n=1}^N \lambda_n}$

- Primal-inner optimal:

$$\frac{\delta L(R, c, \lambda)}{\delta \lambda_n} = 0 = ||z_n - c||^2 - R^2$$

9

By equation 4, if $R > 0$ then $\sum_{n=1}^N \lambda_n = 1$ and thus $c = \frac{\sum_{n=1}^N \lambda_n z_n}{\sum_{n=1}^N \lambda_n} = \sum_{n=1}^N \lambda_n z_n$.

$$\begin{aligned}
L(R, c, \lambda) &= R^2 - R^2 \sum_{n=1}^N \lambda_n + \sum_{n=1}^N \lambda_n (z_n^T z_n - 2z_n^T c + c^T c) \\
&= \sum_{n=1}^N \lambda_n (z_n^T z_n - 2z_n^T c + c^T c) \\
&= \sum_{n=1}^N \lambda_n (z_n^T z_n - 2z_n^T \sum_{i=1}^N \lambda_i z_i + \sum_{i=1}^N \sum_{j=1}^N \lambda_i z_i^T \lambda_j z_j) \\
&= \sum_{n=1}^N \lambda_n z_n^T z_n - 2 \sum_{n=1}^N \lambda_n z_n^T \sum_{i=1}^N \lambda_i z_i + \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j z_i^T z_j \\
&= \sum_{n=1}^N \lambda_n z_n^T z_n - \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j z_i^T z_j
\end{aligned}$$

which $\sum_{n=1}^N \lambda_n = 1$.

10

By the KKT condition of primal-inner optimal $0 = \|z_i - c\|^2 - R^2$.

$$\begin{aligned}
R &= \|z_i - c\| \\
&= (z_i^T z_i - 2z_i^T c + c^T c)^{\frac{1}{2}} \\
&= (z_i^T z_i - 2z_i^T \sum_{n=1}^N \lambda_n z_n + \sum_{i=n}^N \sum_{m=1}^N \lambda_n \lambda_m z_n^T z_m)^{\frac{1}{2}} \\
&= \left(K(x_i, x_i) - 2 \sum_{n=1}^N \lambda_n K(x_i, x_n) + \sum_{n=1}^N \sum_{m=1}^N \lambda_n \lambda_m K(x_n, x_m) \right)^{\frac{1}{2}}
\end{aligned}$$

11

The dual problem of hard-margin and soft-margin SVM are almost the same. The only difference is soft-margin has one more constraint which is $\alpha_n \leq C$ for $n = 1, \dots, N$. Therefore, when the optimal α_n^* in the hard-margin SVM satisfy $C \geq \max_{1 \leq n \leq N} \alpha_n^*$ which will also satisfy $\alpha_n \leq C$ for $n = 1, \dots, N$ and the α_n^* will also be the optimal solution of soft-margin SVM.

12

The original minimization objective of soft-margin SVM is

$$\frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} - \mathbb{1}^T \boldsymbol{\alpha} \quad (3)$$

Let the new objective of soft-margin with new kernel $\tilde{K}(x, x') = pK(x, x')$ and new $\tilde{C} = \frac{C}{p}$ be

$$\begin{aligned} \frac{1}{2}\boldsymbol{\alpha}'^T Q' \boldsymbol{\alpha}' - \mathbb{1}^T \boldsymbol{\alpha}' &= \frac{1}{2}\boldsymbol{\alpha}'^T pQ \boldsymbol{\alpha}' - \mathbb{1}^T \boldsymbol{\alpha}' \\ &= \frac{1}{p} \left(\frac{1}{2}(p\boldsymbol{\alpha}')^T Q (p\boldsymbol{\alpha}') - \mathbb{1}^T (p\boldsymbol{\alpha}') \right) \end{aligned} \quad (4)$$

Since $\frac{1}{p}$ constant in (4) can be ignored, the objective function of (3) is equivalent to (4) in the case of $\boldsymbol{\alpha}' = \frac{\boldsymbol{\alpha}}{p}$. And $\boldsymbol{\alpha}'$ also satisfies the condition of C' .

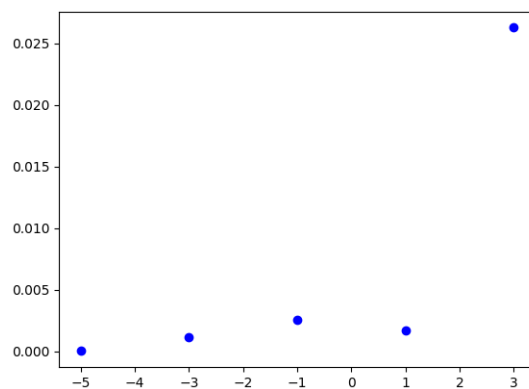
$$0 \leq \alpha' = \frac{\alpha}{p} \leq C' = \frac{C}{p}$$

The new decision function $f(x)$ is equivalent to the original decision function $f(x)$.

$$\begin{aligned} f'(x) &= \sum_{\text{SV indices } n} \alpha'_n y_n \tilde{K}(x_n, x) + y_s - \sum_{\text{SV indices } n} \alpha'_n y_n \tilde{K}(x_n, x_s) \\ &= \sum_{\text{SV indices } n} \alpha_n y_n K(x_n, x) + y_s - \sum_{\text{SV indices } n} \alpha_n y_n K(x_n, x_s) \\ &= f(x) \end{aligned}$$

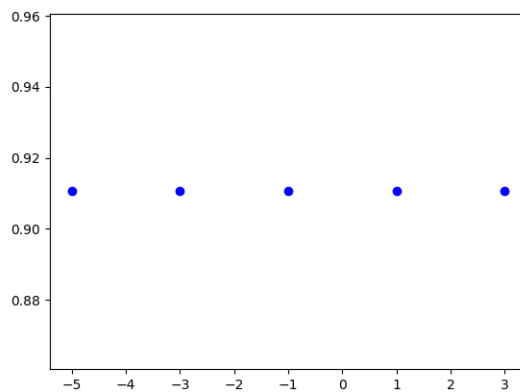
which (x_s, y_s) is a free support vector.

13



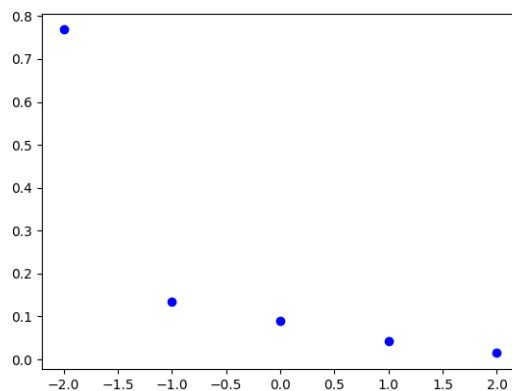
$\|w\|$ has the largest value when $\log_{10} C = 3$ and stays the same on the rest four C .

14



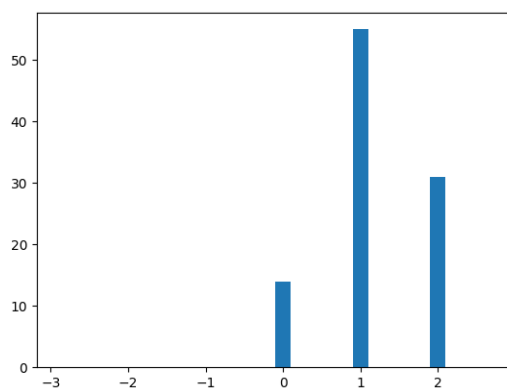
E_{in} remains the same for each C . Which means that the SVM models with different C have the same performance on training samples.

15



The distance between free support vectors and hyperplane has the largest save when $C = 0.01$. This verifies the intuition of soft-margin SVM objective function, smaller C means smaller penalty and larger margin.

16



Only the models with $\gamma = 1, 10, 100$ are selected. Through validation, $\gamma = 10$ seems to have best performance.

17

In dual SVM KKT condition, we know that

$$w_i = \sum_{n=1}^N \alpha_n y_n z_{n,i} \quad (5)$$

and

$$\sum_{n=1}^N \alpha_n y_n = 0 \quad (6)$$

$$\begin{aligned} w_0 &= \sum_{n=1}^N \alpha_n y_n z_{n,0} \\ &= z_{n,0} \sum_{n=1}^N \alpha_n y_n = 0 \end{aligned}$$

18

Assume that the new objective function is

$$\frac{1}{2} \boldsymbol{\alpha}'^T Q' \boldsymbol{\alpha}' - \mathbb{1}^T \boldsymbol{\alpha}'$$

which $Q' = Q + q \cdot \mathbf{y} \mathbf{y}^T$ because of

$$Q'_{n,m} = y_n y_m \tilde{K}(x_n, y_n) = y_n y_m K(x_n, y_n) + q y_n y_m$$

We also know that $\sum_{i=1}^N \alpha'_i y_i = 0$ and so $\mathbf{y}^T \boldsymbol{\alpha}' = 0$.

$$\begin{aligned} \frac{1}{2} \boldsymbol{\alpha}'^T Q' \boldsymbol{\alpha}' - \mathbb{1}^T \boldsymbol{\alpha}' &= \frac{1}{2} \boldsymbol{\alpha}'^T (Q + q \mathbf{y} \mathbf{y}^T) \boldsymbol{\alpha}' - \mathbb{1}^T \boldsymbol{\alpha}' \\ &= \frac{1}{2} \boldsymbol{\alpha}'^T Q \boldsymbol{\alpha}' + \frac{1}{2} (\mathbf{y}^T \boldsymbol{\alpha}') (\mathbf{y}^T \boldsymbol{\alpha}') - \mathbb{1}^T \boldsymbol{\alpha}' \\ &= \frac{1}{2} \boldsymbol{\alpha}'^T Q \boldsymbol{\alpha}' - \mathbb{1}^T \boldsymbol{\alpha}' \end{aligned}$$

The new objective function is equivalent to the original one. Therefore, the support vectors for each SVM are the same and $\alpha'_i = \alpha_i$. The new decision function $f'(x)$ is the same as $f(x)$.

$$\begin{aligned}
f'(x) &= \sum_{\text{SV indices } n} \alpha'_n y_n \tilde{K}(x_n, x) + y_s - \sum_{\text{SV indices } n} \alpha'_n y_n \tilde{K}(x_n, x_s) \\
&= \left(\sum_{\text{SV indices } n} \alpha'_n y_n K(x_n, x) + q \sum_{\text{SV indices } n} \alpha'_n y_n \right) \\
&\quad + y_s - \left(\sum_{\text{SV indices } n} \alpha'_n y_n K(x_n, x_s) + q \sum_{\text{SV indices } n} \alpha'_n y_n \right) \\
&= \sum_{\text{SV indices } n} \alpha_n y_n K(x_n, x) + y_s - \sum_{\text{SV indices } n} \alpha_n y_n K(x_n, x_s) \\
&= f(x)
\end{aligned}$$

which (x_s, y_s) is a free support vector.