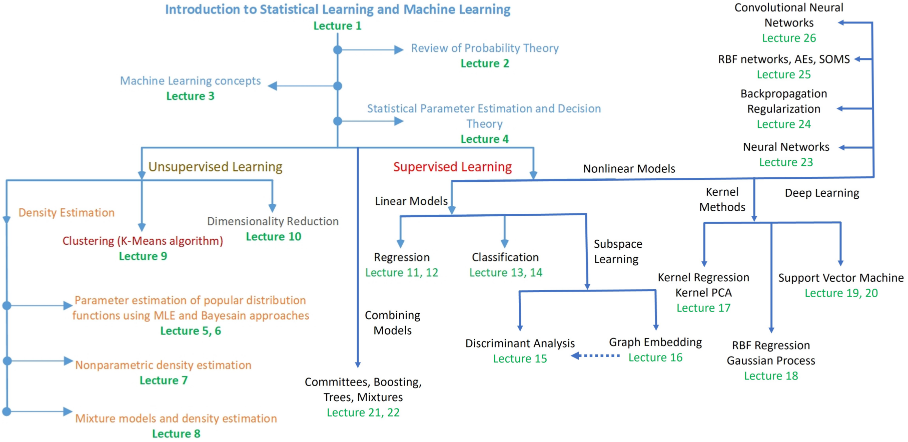
Statistical Learning and Machine Learning

Lecture 20 - Sparse Kernel Machines

October 13, 2021

Course overview and where do we stand



Support Vector Machine: Linearly separable case

Binary linear decision function defined on $\phi(\mathrm{x})$:

$$y(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x}) + b \tag{1}$$

where b is the bias (off-set from the origin) parameter.

Given a set of data points $\mathrm{x}_n,\ n = 1, \ldots, N$, the corresponding class labels $t_n \in \{-1, 1\}$ and parameters w and b classifying correctly all data points:

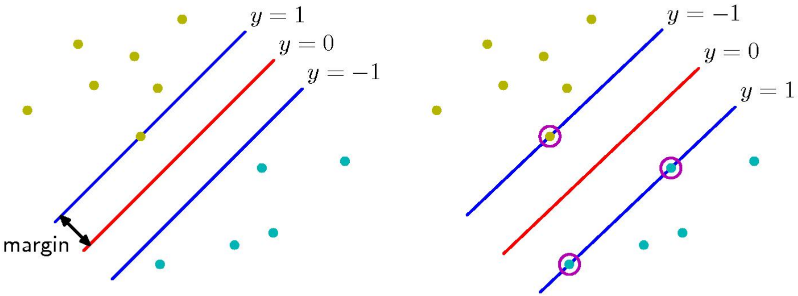
- For $\mathrm{x}_n \in \mathcal{C}_1$: $y(\mathrm{x}) > 0$
- For $\mathrm{x}_n \in \mathcal{C}_2$: $y(\mathrm{x}) < 0$
- We can unify the two above cases using:

$$t_n y(\mathrm{x}) > 0, \ \ n = 1, \ldots, N. \tag{2}$$

For linearly separable classes, there may be multiple combinations of w and b satisfying the above conditions.

Support Vector Machine: Linearly separable case

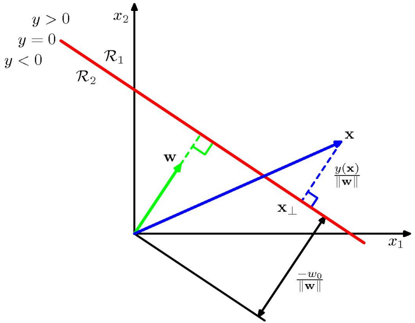
SVM selects the w and b maximizing the *margin* between the two classes.



The margin is determined by a subset of the data points, known as *support vectors*. SVM defines the decision function using only the support vectors (*sparse kernel machine*).

Support Vector Machine: Linearly separable case

Reminder: The perpendicular distance of a point x from a hyperplane defined by $y(\mathrm{x}) = 0$ is given by $|y(\mathrm{x})|/\|\mathrm{w}\|$.



Thus the distance of a point x_n to the decision hyperplane is:

$$\frac{t_n y(\mathrm{x}_n)}{\|\mathrm{w}\|} = \frac{t_n(\mathrm{w}^T\phi(\mathrm{x}_n) + b)}{\|\mathrm{w}\|} \tag{3}$$

Support Vector Machine: Linearly separable case

The margin is defined by the perpendicular distance to the closest point x_n. Thus:

$$\arg\max_{\mathrm{w},b} \left\{ \frac{1}{\|\mathrm{w}\|} \min_n \left[t_n(\mathrm{w}^T \phi(\mathrm{x}_n) + b) \right] \right\} \tag{4}$$

To solve the above problem, we use the following trick:

- We observe that by rescaling $\mathrm{w} \to k\mathrm{w}$ and $b \to kb$ the decision function $t_n y(\mathrm{x}_n)/\|\mathrm{w}\|$ does not change
- we (implicitly) use a k such that for the closest data point to the decision hyperplane x_j:

$$t_j \left(\mathrm{w}^T \phi(\mathrm{x}_j) + b \right) = 1 \tag{5}$$

- then for all data points:

$$t_n \left(\mathrm{w}^T \phi(\mathrm{x}_n) + b \right) \geq 1, \;\; n = 1, \ldots, N \tag{6}$$

Support Vector Machine: Linearly separable case

The decision function maximizing the margin ($\|\mathrm{w}\|^{-1}$) can be obtained by optimizing for:

$$\arg\min_{\mathrm{w},b} \frac{1}{2}\|\mathrm{w}\|^2 \tag{7}$$

subject to the constraints:

$$t_n\left(\mathrm{w}^T\phi(\mathrm{x}_n) + b\right) \geq 1, \;\; n = 1, \ldots, N \tag{8}$$

To solve the problem above, we introduce Lagrange multipliers $\alpha_n \geq 0$:

$$L(\mathrm{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{n=1}^{N} \alpha_n \left[t_n\left(\mathrm{w}^T\phi(\mathrm{x}_n) + b\right) - 1\right] \tag{9}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^T$.

Support Vector Machine: Linearly separable case

Setting the derivatives of $L(\mathbf{w}, b, \boldsymbol{\alpha})$ w.r.t. $\mathbf{w}$ and b to zero we obtain:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \phi(\mathbf{x}_n) \tag{10}$$

$$0 = \sum_{n=1}^{N} \alpha_n t_n \tag{11}$$

Eliminating $\mathbf{w}$ and b from $L(\mathbf{w}, b, \boldsymbol{\alpha})$ gives the *dual representation* of the problem, in which we maximize:

$$\tilde{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \kappa(\mathbf{x}_n, \mathbf{x}_m) \tag{12}$$

subject to constraints $\sum_n \alpha_n t_n = 0$ and $\alpha_n \geq 0,\ n = 1, \ldots, N$.
The kernel is $\kappa(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$.

Support Vector Machine: Linearly separable case

After solving for $\tilde{L}(\boldsymbol{\alpha})$ (quadratic problem w.r.t. $\boldsymbol{\alpha}$), b is given by:

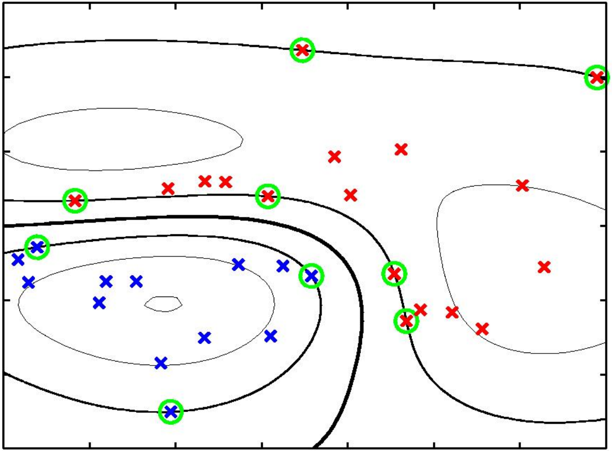
$$b = \frac{1}{N_{\mathcal{S}}} \sum_{n \in \mathcal{S}} \left(t_n - \sum_{m \in \mathcal{S}} \alpha_m t_m \kappa(\mathrm{x}_n, \mathrm{x}_m) \right) \tag{13}$$

where $\mathcal{S}$ is the index set of support vectors and $N_{\mathcal{S}}$ is the number of support vectors.

To classify a new data point x_* we evaluate the sign of:

$$y(\mathrm{x}_*) = \sum_{n=1}^{N} \alpha_n t_n \kappa(\mathrm{x}_n, \mathrm{x}_*) + b. \tag{14}$$

Support Vector Machine: Non-linearly separable case



Contours of constant $y(\mathrm{x})$ of an SVM using RBF kernel function

Support Vector Machine: Non-linearly separable case

The decision function maximizing the margin and allowing some data points to be mis-classified can be obtained by optimizing for:

$$\arg\min_{\mathrm{w},b} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{n=1}^{N}\xi_n \tag{15}$$

subject to the constraints:

$$\begin{aligned} t_n\left(\mathrm{w}^T\phi(\mathrm{x}_n)+b\right) &\geq 1-\xi_n,\;\; n=1,\ldots,N & (16)\\ \xi_n &\geq 0,\;\; n=1,\ldots,N & (17)\end{aligned}$$

where $\xi_n,\; n=1,\ldots,N$ are the *slack variables* and $C>0$ is a hyper-parameter controlling the importance of the two terms.

For a correctly classified data point x_j the corresponding slack variable is $\xi_j=0$.

Support Vector Machine: Non-linearly separable case

To solve the problem above, we introduce Lagrange multipliers $\alpha_n \geq 0$ and $\mu_n \geq 0$:

$$L(\mathrm{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}\alpha_n\left[t_n\left(\mathrm{w}^T\phi(\mathrm{x}_n) + b\right) - 1 + \xi_n\right] - \sum_{n=1}^{N}\mu_n\xi_n \tag{18}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]$.

Setting the derivatives of $L(\mathrm{w}, b, \boldsymbol{\alpha})$ w.r.t. w, b and ξ_n, $n = 1, \ldots, N$:

$$\frac{\theta L}{\theta \mathrm{w}} = 0 \quad \Rightarrow \quad \mathrm{w} = \sum_{n=1}^{N}\alpha_n t_n \phi(\mathrm{x}_n) \tag{19}$$

$$\frac{\theta L}{\theta b} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N}\alpha_n t_n = 0 \tag{20}$$

$$\frac{\theta L}{\theta \xi_n} = 0 \quad \Rightarrow \quad \alpha_n = C - \mu_n \tag{21}$$

Support Vector Machine: Non-linearly separable case

We introduce the results obtained by the derivatives above in $L(\mathrm{w}, b, \boldsymbol{\alpha})$ and we obtain the dual Lagrangian:

$$\tilde{L}(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m t_n t_m \kappa(\mathrm{x}_n, \mathrm{x}_m) \tag{22}$$

subject to the constraints $\sum_n \alpha_n t_n = 0$ and $0 \leq \alpha_n \leq C, \; n = 1, \ldots, N$ (*box constraints*).

After solving for $\tilde{L}(\boldsymbol{\alpha})$ (quadratic problem w.r.t. $\boldsymbol{\alpha}$), b is given by:

$$b = \frac{1}{N_{\mathcal{M}}} \sum_{n \in \mathcal{M}} \left(t_n - \sum_{m \in \mathcal{S}} \alpha_m t_m \kappa(\mathrm{x}_n, \mathrm{x}_m) \right) \tag{23}$$

where $\mathcal{S}$ is the set of support vectors, $\mathcal{M}$ is the set of data satisfying the box constraints and $N_{\mathcal{S}}$, $N_{\mathcal{M}}$ are the corresponding set sizes.

Support Vector Machine: Non-linearly separable case

To classify a new data point x_* we evaluate the sign of:

$$y(\mathrm{x}_*) = \sum_{n=1}^{N} \alpha_n t_n \kappa(\mathrm{x}_n, \mathrm{x}_*) + b. \tag{24}$$

Support Vector Machine: $K > 2$

For $K > 2$, we can formulate K we can define an optimization problem optimizing K decision functions of the form:

$$y_k(\mathrm{x}) = \mathrm{w}_k^T \mathrm{x} + b_{k0} \tag{25}$$

and then classify a new sample x_* using:

$$y(\mathrm{x}_*) = \max_k y_k(\mathrm{x}_*) \tag{26}$$

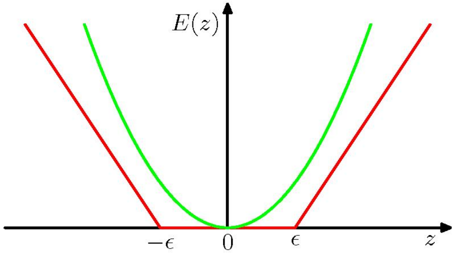
Otherwise, we can combine binary SVM classifiers using the following combination schemes:

- One-versus-rest: Define K binary SVM classifiers, each discriminating between one class the the rest
- One-versus-one: Define $K(K-1)/2$ binary SVM classifiers, one per each class pair

Support Vector Regression

To define SVR we use the ϵ-*sensitive error function*:

$$E_\epsilon(y(\mathrm{x}) - t) = \begin{cases} 0, & \text{if } |y(\mathrm{x}) - t| < \epsilon \\ |y(\mathrm{x}) - t| - \epsilon, & \text{otherwise} \end{cases} \tag{27}$$



Red: ϵ-sensitive error function
Green: quadratic error function

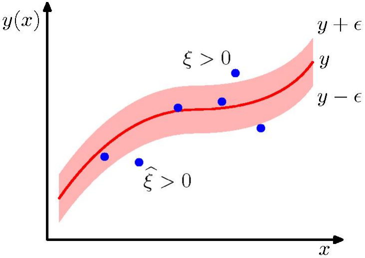
Support Vector Regression

We minimize for:

$$\mathcal{J} = \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{n=1}^{N} E_\epsilon(y(\mathrm{x}_n) - t_n) \tag{28}$$

where $y(\mathrm{x}) = \mathrm{w}^T\phi(\mathrm{x}) + b$.

To account for errors, we can introduce slack variables (we need two for each data point ξ_n and $\hat{\xi}_n$, $n = 1, \ldots, N$).



Support Vector Regression

We minimize for:

$$\mathcal{J} = \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) \tag{29}$$

subject to the constraints:

$$\begin{aligned}
t_n &\leq y(\mathrm{x}_n) + \epsilon + \xi_n, \quad n = 1, \dots, N && (30)\\
t_n &\geq y(\mathrm{x}_n) - \epsilon - \hat{\xi}_n, \quad n = 1, \dots, N && (31)\\
\xi_n &\geq 0, \quad n = 1, \dots, N && (32)\\
\hat{\xi}_n &\geq 0, \quad n = 1, \dots, N && (33)
\end{aligned}$$

Support Vector Regression

We introduce Lagrange multipliers $\alpha_n \geq 0$, $\hat{\alpha}_n \geq 0$, $\mu_n \geq 0$ and $\hat{\mu}_n \geq 0$:

$$\begin{aligned} L &= \frac{1}{2}\|\mathrm{w}\|^2 + C\sum_{n=1}^{N}(\xi_n + \hat{\xi}_n) - \sum_{n=1}^{N}(\mu_n\xi_n + \hat{\mu}_n\hat{\xi}_n) \\ &\quad - \sum_{n=1}^{N}\alpha_n(\epsilon + \xi_n + y(\mathrm{x}_n) - t_n) \\ &\quad - \sum_{n=1}^{N}\hat{\alpha}_n(\epsilon + \hat{\xi}_n - y(\mathrm{x}_n) + t_n) \end{aligned} \tag{34}$$

Support Vector Regression

Setting the derivatives of L w.r.t. w, b, ξ_n and $\hat{\xi}_n$, $n = 1, \ldots, N$ equal to zeros:

$$
\begin{aligned}
\frac{\theta L}{\theta \mathrm{w}} = 0 \quad &\Rightarrow \quad \mathrm{w} = \sum_{n=1}^{N} (\alpha_n - \hat{\alpha}_n) \phi(\mathrm{x}_n) && (35) \\
\frac{\theta L}{\theta b} = 0 \quad &\Rightarrow \quad \sum_{n=1}^{N} (\alpha_n - \hat{\alpha}_n) = 0 && (36) \\
\frac{\theta L}{\theta \xi_n} = 0 \quad &\Rightarrow \quad \alpha_n + \mu_n = C && (37) \\
\frac{\theta L}{\theta \hat{\xi}_n} = 0 \quad &\Rightarrow \quad \hat{\alpha}_n + \hat{\mu}_n = C && (38)
\end{aligned}
$$

Support Vector Regression

Eliminating w, b, ξ_n and $\hat{\xi}_n$, $n = 1, \ldots, N$ from L we get the dual problem maximizing:

$$
\begin{aligned}
\tilde{L}(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}) &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(\alpha_n - \hat{\alpha}_n)(\alpha_m - \hat{\alpha}_m)\kappa(\mathrm{x}_n, \mathrm{x}_m) \\
&\quad -\epsilon \sum_{n=1}^{N}(\alpha_n + \hat{\alpha}_n) + \sum_{n=1}^{N}(\alpha_n - \hat{\alpha}_n)t_n \qquad (39)
\end{aligned}
$$

subject to the constraints $0 \leq \alpha_n \leq C$, $0 \leq \hat{\alpha}_n \leq C$ and $\sum_n(\alpha_n - \hat{\alpha}_n) = 0$.

The parameter b is calculated by:

$$
b = \frac{1}{N_{\mathcal{M}}}\sum_{n \in \mathcal{M}}\left(t_n - \epsilon - \sum_{m=1}^{N}(\alpha_m - \hat{\alpha}_m)\kappa(\mathrm{x}_n, \mathrm{x}_m)\right) \qquad (40)
$$

where $\mathcal{M}$ is the set of data points for which $0 < \alpha_n < C$ or $0 < \hat{\alpha}_n < C$ and $N_{\mathcal{M}}$ is the size of $\mathcal{M}$.

Support Vector Regression

A new data point x_* is evaluated by:

$$y(\mathrm{x}_*) = \sum_{n=1}^{N} (\alpha_n - \hat{\alpha}_n)\kappa(\mathrm{x}_n, \mathrm{x}_*) + b. \tag{41}$$