Support Vector Machines

AW

Lecture Overview

1. Recap
2. SVM
3. Soft Margin SVM

Margins

Concept of a margin: The margin of a hyperplane with respect to a training set = 2$\times$ minimal distance between a point in the training set and the hyperplane.

Definition

D is a distribution over $\mathbb{R}^d \times \{\pm 1\}$. We say that D is separable with a (γ, ρ)-margin if there exists hyperplane $[\overline{w}^* : b^*]$ with $\|w\| = 1$ such that with probability 1 over the choice of $(\overline{x}, y) \sim D$ we have that $y(\overline{w}^* \bullet \overline{x} + b^*) \geq \gamma$ and $\|\overline{x}\| \leq \rho$.

Margins

Concept of a margin: The margin of a hyperplane with respect to a training set = 2$\times$ minimal distance between a point in the training set and the hyperplane.

Definition

D is a distribution over $\mathbb{R}^d \times \{\pm 1\}$. We say that D is separable with a (γ, ρ)-margin if there exists hyperplane $[\overline{w}^* : b^*]$ with $\|w\| = 1$ such that with probability 1 over the choice of $(\overline{x}, y) \sim D$ we have that $y(\overline{w}^* \bullet \overline{x} + b^*) \geq \gamma$ and $\|\overline{x}\| \leq \rho$.

Theorem (Optimal Hyperplane)

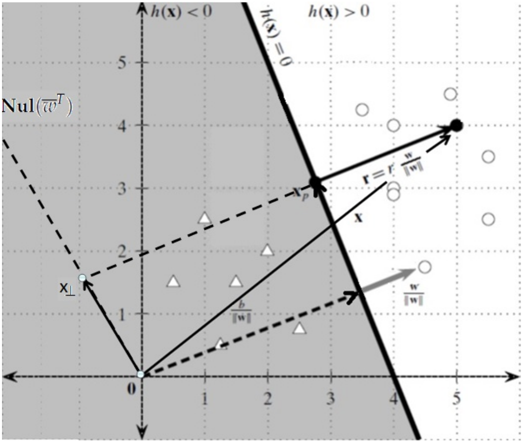
D is a distribution over $\mathbb{R}^d \times \{\pm 1\}$ that is separable with a (γ, ρ)-margin. Then, with probability of at least $1 - \delta$ over the choice of a training set of size m, there is a hyperplane that has upper bound on error $\sqrt{\frac{(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$

Our goal is to find this hyperplane.

Distance of a Point to the Hyperplane

What is the distance of $\overline{x}$ from h along $\overline{w}$? It is the difference of the following terms:

- The length of projection of $\overline{x}$ onto $\overline{w}$ where the projection vector is $\frac{\overline{x}\bullet\overline{w}}{\|\overline{w}\|^2}\overline{w} = \frac{\overline{x}\bullet\overline{w}}{\|\overline{w}\|}\hat{w}$ so length is $\frac{\overline{x}\bullet\overline{w}}{\|\overline{w}\|}$
- The length of translation vector$= \frac{-b}{\|\overline{w}\|}$



This value is negative when $\overline{w} \bullet \overline{x} < -b$ (class -1) and positive when $\overline{w} \bullet \overline{x} > -b$ (class 1). But distance is always positive so we need to take absolute value

$$d(\overline{x}, h) = \left| \frac{\overline{x}\bullet\overline{w}}{\|\overline{w}\|} - \frac{-b}{\|\overline{w}\|} \right| = \frac{y(\overline{x}\bullet\overline{w}+b)}{\|\overline{w}\|}$$

Support Vectors

Distance example: Suppose separating plane is $h = \left[\begin{pmatrix} 4 \\ 3 \end{pmatrix}^T : 20 \right]$.

Then for a point $\overline{x} = \left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, 1 \right)$ distance from h is

$$d(\overline{x}, h) = \frac{1 \left([4 \quad 3] \begin{pmatrix} 3 \\ 4 \end{pmatrix} - 20 \right)}{\sqrt{4^2+3^2}} = \frac{4}{5}$$

Over all the m points in training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we define the margin d^* of the linear classifier as the minimum distance of a point in S from the separating hyperplane h,i.e.

$$d^* = \min_{\overline{x}_i \in S} \left\{ \frac{y_i(\overline{x}_i \bullet \overline{w} + b)}{\|\overline{w}\|} \right\}$$

Note that $d^* \neq 0$, since h is a (strictly) separating hyperplane.
All the vectors in S that are at margin distance d^* from h are the support vectors for the hyperplane h.

Support Vectors

Over all the m points in training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we define the margin d^* of the linear classifier as the minimum distance of a point in S from the separating hyperplane h,i.e.

$$d^* = \min_{\overline{x}_i \in S} \left\{ \frac{y_i(\overline{x}_i \bullet \overline{w} + b)}{\|\overline{w}\|} \right\}$$

Note that $d^* \neq 0$, since h is a (strictly) separating hyperplane.
All the vectors in S that are at margin distance d^* from h are the support vectors for the hyperplane h.

Example continued: Let $S =$
$\left\{ \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 2.5 \\ 2 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, 1 \right); \left(\begin{pmatrix} 1.5 \\ 6 \end{pmatrix}, 1 \right) \right\}$

We can check that $h = \left[\begin{pmatrix} 4 \\ 3 \end{pmatrix}^T : 20 \right]$ is a separating hyperplane for S.What are suport vectors of this hyperplane?

Support Vectors

Over all the m points in training set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we define the margin d^* of the linear classifier as the minimum distance of a point in S from the separating hyperplane h,i.e.

$$d^* = \min_{\overline{x}_i \in S} \left\{ \frac{y_i(\overline{x}_i \bullet \overline{w} + b)}{\|\overline{w}\|} \right\}$$

Note that $d^* \neq 0$, since h is a (strictly) separating hyperplane.
All the vectors in S that are at margin distance d^* from h are the support vectors for the hyperplane h.

Example continued: Let $S =$

$$\left\{ \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 2 \\ 1 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 2.5 \\ 2 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, 1 \right); \left(\begin{pmatrix} 1.5 \\ 6 \end{pmatrix}, 1 \right) \right\}$$

We can check that $h = \left[\begin{pmatrix} 4 \\ 3 \end{pmatrix}^T : 20 \right]$ is a separating hyperplane for S.

Easy to check that $d^* = \frac{4}{5}$ for h and that support vectors are $\left(\begin{pmatrix} 2.5 \\ 2 \end{pmatrix}, -1 \right)$, $\left(\begin{pmatrix} 1.5 \\ 6 \end{pmatrix}, 1 \right)$ and $\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, 1 \right)$.

Lecture Overview

Canonical Hyperplane

- For hyperplane $h = [\overline{w} : b]$ multiplying on both sides by a scalar s yields an equivalent hyperplane representation $h = [s\overline{w} : sb]$. To obtain the canonical hyperplane, set s so that $d^* \|\overline{w}\| = 1$.
- So for any support vector $(\overline{x}^*, y^*)$ holds $sy^*(\overline{w} \bullet \overline{x}^* + b) = 1$. Thus $s = \frac{1}{y^*(\overline{w} \bullet \overline{x}^* + b)}$.

Example continued: For all support vectors $\left(\begin{pmatrix} 2.5 \\ 2 \end{pmatrix}, -1\right)$, $\left(\begin{pmatrix} 1.5 \\ 6 \end{pmatrix}, 1\right)$ and $\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, 1\right)$ of the hyperplane $h = \left[\begin{pmatrix} 4 \\ 3 \end{pmatrix}^T : 20\right]$ holds $s = \frac{1}{y^*(\overline{w} \bullet \overline{x}^* + b)} = \frac{1}{4}$. So canonical representation of the hyperplane is $h = \left[\begin{pmatrix} 1 \\ 3/4 \end{pmatrix}^T : 5\right]$ and $d^* = \frac{1}{5}$.

Canonical Hyperplane

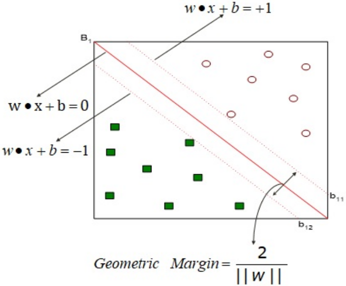
- For hyperplane $h = [\overline{w} : b]$ multiplying on both sides by a scalar s yields an equivalent hyperplane representation $h = [s\overline{w} : sb]$. To obtain the canonical hyperplane, set s so that $d^* \|\overline{w}\| = 1$.
- So for any support vector $(\overline{x}^*, y^*)$ holds $sy^*(\overline{w} \bullet \overline{x}^* + b) = 1$. Thus $s = \frac{1}{y^*(\overline{w} \bullet \overline{x}^* + b)}$.
- Given the canonical hyperplane $h = [\overline{w} : b]$
  - for each support vector $(\overline{x}^*, y^*)$ holds $y^*(\overline{w} \bullet \overline{x}^* + b) = 1$
  - for any point $(\overline{x}_i, y_i) \in S$ that is not a support vector we have $y_i(\overline{w} \bullet \overline{x}_i + b) > 1$, because, it is farther from the hyperplane than a support vector

  Thus for all $(\overline{x}_i, y_i) \in S$ holds $y_i(\overline{w} \bullet \overline{x}_i + b) \geq 1$

Maximum Margin Hyperplane

Let $\mathscr{H}$ be class of linear hyperplanes. For canonical hyperplane h distance of a support vector from hyperplane is $d^* = \frac{1}{\|\overline{w}\|}$, so geometric margin of a hyperplane is at least $2d^* = \frac{2}{\|\overline{w}\|}$.
So we need

$$h^* = \arg\max_{h_S \in \mathscr{H}} \{d^*_{h_S}\} = \arg\max_{\overline{w},b} \left\{ \frac{2}{\|\overline{w}\|} \right\}$$



under constraints $y_i(\overline{w} \bullet \overline{x}_i + b) \geq 1$ for all $(\overline{x}_i, y_i) \in S$.

Instead of maximizing margin we can minimize its inverse, and then instead of norm (to get rid of square roots) we can minimize the square of norm, the resulting $\arg\min$ will be the same as original $\arg\max$. So we need to solve

$$\begin{array}{ll} & \min_{\overline{w},b} \left\{ \frac{\|\overline{w}\|^2}{2} \right\} \\ \text{subject to} & y_i(\overline{w} \bullet \overline{x}_i + b) \geq 1 \quad \text{for all } (\overline{x}_i, y_i) \in S \end{array}$$

Solving SVM Optimization Problem

As before introducing Lagrange multipliers λ_i for each constraint the objective function becomes:

$$\min_{\overline{w},b} L = \min_{\overline{w},b} \left\{ \frac{\|\overline{w}\|^2}{2} - \sum_{i=1}^{m} \lambda_i (y_i(\overline{w} \bullet \overline{x}_i + b) - 1) \right\}$$

We need stationary point solution that also satisfy KKT (Karush-Kuhn-Tacker) conditions (as constraints are inequalities). Simply put we need $\frac{\partial L}{\partial \overline{w}} = 0$ and $\frac{\partial L}{\partial b} = 0$ at the same time $\lambda_i(y_i(\overline{w} \bullet \overline{x}_i + b) - 1) = 0$ and $\lambda_i \geq 0$ for all $0 \leq i \leq m$.

$$\frac{\partial L}{\partial \overline{w}} = \overline{w} - \sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i = 0 \quad \text{or} \quad \overline{w} = \sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i$$

and

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} \lambda_i \, y_i = 0$$

Notice that

$$\sum_{i=1}^{m} \lambda_i (y_i(\overline{w} \bullet \overline{x}_i + b) - 1) = \overline{w} \bullet \underbrace{\sum_{i=1}^{m} \lambda_i y_i \overline{x}_i}_{\overline{w}} + b \underbrace{\sum_{i=1}^{m} \lambda_i y_i}_{0} - \sum_{i=1}^{m} \lambda_i$$

Solving SVM Optimization Problem

As before introducing Lagrange multipliers λ_i for each constraint the objective function becomes:

$$\min_{\overline{w},b} L = \min_{\overline{w},b} \left\{ \frac{\|\overline{w}\|^2}{2} - \sum_{i=1}^{m} \lambda_i (y_i(\overline{w} \bullet \overline{x}_i + b) - 1) \right\}$$

We need stationary point solution that also satisfy KKT (Karush-Kuhn-Tacker) conditions (as constraints are inequalities). Simply put we need $\frac{\partial L}{\partial \overline{w}} = 0$ and $\frac{\partial L}{\partial b} = 0$ at the same time $\lambda_i(y_i(\overline{w} \bullet \overline{x}_i + b) - 1) = 0$ and $\lambda_i \geq 0$ for all $0 \leq i \leq m$.

$$\frac{\partial L}{\partial \overline{w}} = \overline{w} - \textstyle\sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i = 0 \quad \text{or} \quad \overline{w} = \textstyle\sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i$$

and

$$\frac{\partial L}{\partial b} = \textstyle\sum_{i=1}^{m} \lambda_i \, y_i = 0$$

Notice that $\sum_{i=1}^{m} \lambda_i(y_i(\overline{w} \bullet \overline{x}_i + b) - 1) = \overline{w} \bullet \overline{w} - \sum_{i=1}^{m} \lambda_i$

Solving SVM Optimization Problem

As before introducing Lagrange multipliers λ_i for each constraint the objective function becomes:

$$\min_{\overline{w},b} L = \min_{\overline{w},b} \left\{ \frac{\|\overline{w}\|^2}{2} - \sum_{i=1}^{m} \lambda_i (y_i(\overline{w} \bullet \overline{x}_i + b) - 1) \right\}$$

We need stationary point solution that also satisfy KKT (Karush-Kuhn-Tacker) conditions (as constraints are inequalities). Simply put we need $\frac{\partial L}{\partial \overline{w}} = 0$ and $\frac{\partial L}{\partial b} = 0$ at the same time $\lambda_i(y_i(\overline{w} \bullet \overline{x}_i + b) - 1) = 0$ and $\lambda_i \geq 0$ for all $0 \leq i \leq m$.

$$\frac{\partial L}{\partial \overline{w}} = \overline{w} - \textstyle\sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i = 0 \quad \text{or} \quad \overline{w} = \textstyle\sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i$$

and

$$\frac{\partial L}{\partial b} = \textstyle\sum_{i=1}^{m} \lambda_i \, y_i = 0$$

So when pugged into $\min L$ becomes

$\min_{\overline{w},b} L = \min_{\overline{w},b} \left\{ \frac{\|\overline{w}\|^2}{2} - \|\overline{w}\|^2 + \sum_{i=1}^{m} \lambda_i \right\}$

Solving SVM - continued

So we have

$$\min_{\overline{w},b} L = \min_{\overline{w},b} \left\{ \frac{-\|\overline{w}\|^2}{2} + \sum_{i=1}^{m} \lambda_i \right\}$$

where $\overline{w} = \sum_{i=1}^{m} \lambda_i\, y_i\, \overline{x}_i$. Thus substituting, and noticing that $\max_{\overline{\lambda}}$ under constraints $\frac{\partial L}{\partial b} = \sum_{i=1}^{m} \lambda_i\, y_i = 0$ for all i, gives us $\min_{\overline{w},b}$ under constraints $y_i(\overline{w} \bullet \overline{x}_i + b) \geq 1$ for all i we obtain dual quadratic program with *linear* constraints

$$\min_{\overline{w},b} L \quad = \quad \max_{\overline{\lambda}} \left\{ \sum_{i=1}^{m} \lambda_i - \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \overline{x}_i^T \overline{x}_j}{2} \right\}$$

subject to

$$\sum_{i=1}^{m} \lambda_i\, y_i = 0$$
$$\lambda_i \geq 0 \qquad \text{for } 1 \leq i \leq m$$

where $\overline{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$.

This program can be solved by standard methods (e.g. gradient ascent), which gives optimal values $\lambda_1, \ldots, \lambda_m$.

Finding $\overline{w}$ and b

By KKT conditions we have that for all λ_i holds $\lambda_i(y_i(\overline{w} \bullet \overline{x}_i + b) - 1) = 0$ one of the two possibilities hold

- $\lambda_i > 0$ and then $y_i(\overline{w} \bullet \overline{x}_i + b) = 1$, i.e. $\overline{x}_i$ is a support vector
- $\lambda_i = 0$ and $y_i(\overline{w} \bullet \overline{x}_i + b) > 1$, i.e. $\overline{x}_i$ is not a support vector

In other words, since $\overline{w} = \sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i$, orthogonal vector of a hyperplane is a combination of support vectors. It is computed as

$$\overline{w} = \sum_{i:\lambda_i > 0} \lambda_i y_i \, \overline{x}_i$$

Now as before we can find projections of each support vector $\overline{x}_i$ onto $\overline{w}$ and average these projections to get a separation point, but rather than normalizing $\overline{w}$, we can compute these from $y_i(\overline{w} \bullet \overline{x}_i + b) = 1$ as $b_i = y_i - \overline{w} \bullet \overline{x}_i$ and then take average

$$b = \frac{1}{|\{i : \lambda_i > 0\}|} \sum_{\lambda_i > 0} (y_i - \overline{w} \bullet \overline{x}_i) \, .$$

Error again

Since margins are the same for all supporting vectors instead of solving $y_i(\overline{w} \bullet \overline{x}_i + b) = 1$ for all support vectors and averaging, we could solve it for one vector $\overline{x}_i$ in class -1 and one vector $\overline{x}_j$ in class 1 and take $b = \frac{b_i + b_j}{2}$.

Let distribution D from be separable with probability 1 over the choice of data $(\overline{x}, y) \sim D$ with some margin γ.

- Margin γ of optimal (over all data) hyperplane is at most the same as learned margin $|b - b_i|$ (where b_i projection coordinate of any support vector)
- Any computed from training set S hyperplane other that SVM-hyperplane $[\overline{w} : b]$ already has smaller margin than $[\overline{w} : b]$

Hyperplanes have finite VC-dimensions, so by Fundamental theorem they are ERM-agnostically learnable. Since D is separable and if $m > m(\epsilon, \delta)$ our hyperplane is optimal. If there is radius ρ for which $\|\overline{x}\| \leq \rho$ for all datapoints $\overline{x}$ then by Optimal Hyperplane theorem $[\overline{w} : b]$ has the error bound $\sqrt{\frac{(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$.

SVM in R

```
library(e1071);library(mlbench)
data(BreastCancer)
BC<-BreastCancer[!rowSums(is.na(BreastCancer)),-1]
#remove id column and rows with na valuesSVM
#can't handle it
dtrain<-sample(1:nrow(BC),2/3*nrow(BC),F)
#randomly select record #'s for training.
# remaining -dtrain numbers used for test
BC_model<- svm(Class ~ .,BC[dtrain,],
                type='C',kernel='linear')
#kernel linear is a vector (line);
#can be another curve
summary(BC_model)
pred<-predict(BC_model,BC[dtrain,-10]).
table(pred, BC[dtrain, ]$Class)
table(predicted=predict(BC_model,BC[-dtrain,-10]),
              true=BC[-dtrain, ]$Class)
```

Lecture Overview

Soft Margins

If data that is linearly separable there is no $\overline{w}, b$ that satisfy constraints of maximization problem! No solution will be returned. How to modify optimization problem to handle inseparable data?

Key Idea: use of slack parameters. *Intuition*: suppose $[\overline{w}, b]$ reliably separates $99.9(9)\%$ of data points with a large margin. But there is couple data points that end up on the wrong side of the hyperplane, perhaps these are noisy.

- Suppose we can 'move' these points across the hyperplane to the correct side
- We pay for the 'move' of a point, small 'cost', so that a total cost is acceptable if we are not moving a lot of points around

Soft Margin SVM explained

- We formalize the notion of moving to the right side by introducing one slack variable for each training example:

$$y_i(\overline{w} \bullet \overline{x}_i + b) + \xi_i \geq 1 \text{ where } \xi_i \geq 0$$

- We formalize the notion of cost by penalizing oneself for having to use slack in the objective. The smaller the slack the less we pay:

$$\min_{\overline{w},b,\overline{\xi}} \left\{ \frac{\|\overline{w}\|^2}{2} + C \sum_{i=1}^{m} (\xi_i)^k \right\}$$

C and k are parameters of Soft margin SVM:

- k defines loss function $f(\overline{\xi}) \sum_{i=1}^{m} (\xi_i)^k$ - total cost that we pay for violating constraints; $k = 1$ called hinge loss, $k = 2$ quadratic loss
- C is a regularization constant that controls the trade-off between maximizing the margin (minimizing $\frac{\|\overline{w}\|^2}{2}$) and minimizing loss (minimizing total error = distance away from hyperplane $\sum_{i=1}^{m} (\xi_i)^k$)

Solving Soft Margin SVM w/Hinge Loss

Soft Margin SVM with hinge loss optimization problem:

$$\begin{array}{lll} & \min_{\overline{w},b,\overline{xi}} \left\{ \frac{\|\overline{w}\|^2}{2} + C \sum_{i=1}^{m} \xi_i \right\} & \\ \text{subject to} & y_i(\overline{w} \bullet \overline{x}_i + b) + \xi_i \geq 1 & \text{for all } (\overline{x}_i, y_i) \in S \\ & \xi_i \geq 0 & \text{for all } 1 \leq i \leq m \end{array}$$

Introducing Lagrange multipliers λ_i for margin constraints and β_i for non–negativity of slack variables constraints we get

$$\min_{\overline{w},b} L = \min_{\overline{w},b,\overline{\xi}} \left\{ \frac{\|\overline{w}\|^2}{2} - \sum_{i=1}^{m} \lambda_i (y_i(\overline{w} \bullet \overline{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{m} (\beta_i - C)\xi_i \right\}$$

with KKT conditions being $\lambda_i(y_i(\overline{w} \bullet \overline{x}_i + b) - 1 + \xi_i) = 0$ and $\beta_i \xi_i = 0$.
Equating partial derivatives wrt $\overline{w}, b$ and all ξ_i to 0 we get:

$$\begin{array}{lll} & \overline{w} = \sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i & \text{as before} \\ & \sum_{i=1}^{m} \lambda_i \, y_i = 0 & \text{as before} \\ \text{and} & & \\ & C - \lambda_i - \beta_i = 0 & \text{for all } 1 \leq i \leq m \end{array}$$

Soft Margin SVM w/Hinge Loss - continued

Substituting

$$\overline{w} = \sum_{i=1}^{m} \lambda_i \, y_i \, \overline{x}_i \quad \text{as before}$$
$$\sum_{i=1}^{m} \lambda_i \, y_i = 0 \quad \text{as before}$$

and

$$C - \lambda_i - \beta_i = 0 \quad \text{for all } 1 \le i \le m$$

back into

$$\min_{\overline{w},b} L = \min_{\overline{w},b,\overline{\xi}} \left\{ \frac{\|\overline{w}\|^2}{2} - \sum_{i=1}^{m} \lambda_i (y_i(\overline{w} \bullet \overline{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{m} (\beta_i - C)\xi_i \right\}$$

we obtain dual program that is almost the same as in separable case except to the range of λ_i

$$\min_{\overline{w},b} L \quad = \quad \max_{\overline{\lambda}} \left\{ \sum_{i=1}^{m} \lambda_i - \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \overline{x}_i^T \overline{x}_j}{2} \right\}$$

subject to

$$\sum_{i=1}^{m} \lambda_i \, y_i = 0$$
$$C \ge \lambda_i \ge 0 \quad \text{for } 1 \le i \le m$$

The range of λ_i is determined by constraint $C = \lambda_i + \beta_i$ and the fact that $\beta_i \ge 0$

Soft Margin SVM w/Hinge Loss - continued

Solving

$$\min_{\overline{w},b} L \quad = \quad \max_{\overline{\lambda}} \left\{ \sum_{i=1}^{m} \lambda_i - \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \overline{x}_i^T \overline{x}_j}{2} \right\}$$

subject to

$$\sum_{i=1}^{m} \lambda_i \, y_i = 0$$

$$C \geq \lambda_i \geq 0 \qquad \text{for } 1 \leq i \leq m$$

gives optimal values $\lambda_1, \ldots, \lambda_m$. As before support vectors are those for which $\lambda_i > 0$. So $\overline{x_i}$ for which holds $y_i(\overline{w} \bullet \overline{x}_i + b) + \xi_i = 1$ are support vectors (due to KKT conditions). They now include all points that are on the margin (which have zero slack $\xi_i = 0$), as well as all points with positive slack ($\xi_i > 0$)!
As before orthogonal vector of the hyperplane is $\overline{w} = \sum_{\lambda_i > 0} \lambda_i y_i \, \overline{x}_i$.
Since $C = \lambda_i + \beta_i$ and $\beta_i \xi_i = 0$ (KKT condition) we get that those $\overline{x}_i$ for which $0 < \lambda_i < C$ are exactly on the margins ($\xi_i = 0$ while those for which $\lambda_i = C$ are behind hyperplane. To compute bias b we take points on the margins and as before compute $b_i = y_i - \overline{w} \bullet \overline{x}_i$ and then take

Reading

TSKK (main textbook) Section 4.9
Zaki, Meira, Sections 21.1, 21.2, 21.3.1.