CSCI 590: Machine Learning

Lecture 13:

VC theory and support vector machines (SVMs)

Instructor: Murat Dundar

Acknowledgements:
1. SVM Tutorial by C. Burges (http://research.microsoft.com/pubs/67119/svmtutorial.pdf)

Support vector machines

Suppose we have a separating hyperplane $w^T x + b = 0$.

Let $d_+(d_-)$ be the shortest distance from the separating hyperplane to the closest positive (negative) example.

We define the "margin" of the separating hyperplane as $d_+ + d_-$

For the linearly separable case, SVM looks for the separating hyperplane with largest margin.

Defining the margin (1)

Let us suppose that all training data satisfy the following

$$w^T x_i + w_0 \geq 1 \text{ for } y_i = 1$$
$$w^T x_i + w_0 \leq -1 \text{ for } y_i = -1$$

More compactly

$$y_i(w^T x_i + w_0) \geq 1 \; \forall i$$

Defining the margin (2)

Pick a point x_1 on $H_+: w^T x_1 + w_0$ and pick another point x_2 perpendicular to x_1 on $H_-: w^T x_2 + w_0$.

Note that w is orthonormal to both H_+ and H_- and thus, the inner product of the vector x_1- x_2 with w is

$$(x_1 - x_2)^T w = \|x_1 - x_2\| \|w\| \cos(0)$$

We know that the perpendicular distance between H_+ and H_- is

$$w^T (x_1 - x_2) = 2$$

Thus, the margin between H_+ and H_- is $\|x_1 - x_2\| = 2/\|w\|$

Linearly separable case (1)

Maximizing the margin $\frac{2}{\|w\|}$ is equivalent to minimizing $\frac{\|w\|}{2}$ or $\frac{\|w\|^2}{2}$

Note that $\frac{2}{\|w\|}$ defines the margin under the assumption that

$$y_i(w^T x_i + w_0) \geq 1 \; \forall i$$

We should minimize $\frac{\|w\|^2}{2}$ subject to these constraints

Linearly separable case (2)

Optimization problem for the linearly separable case:

$$\text{minimize } \frac{\|w\|^2}{2}$$

$$\text{s. t. } y_i(w^T x_i + w_0) \geq 1 \; \forall i$$

This is a quadratic programming problem with inequality constraints.

Linearly separable case (3)

Karush-Kuhn-Tucker Theorem: If x^* is a local minimum to a constrained problem with equality and inequality constraints, then there exist constants μ^* and λ^*, called KKT multipliers, such that

$$\min_x f(x)\ s.t.\ g(x) \geq 0\ \ h(x) = 0$$

$$\nabla f(x^*) - \nabla g(x^*)\mu^* + \nabla h(x^*)\lambda^* = 0$$

$$h(x^*) = 0 \quad g(x^*) \geq 0$$
$$\mu^* \geq 0 \quad \mu^* \mathrm{g}(x^*) = 0$$

Linearly separable case (4)

The Lagrange equation:

$$L_p = \frac{\|w\|^2}{2} - \sum_{i=1}^{N} \mu_i (y_i (w^T x_i + w_0) - 1)$$

KKT conditions:

1. $w^* - \sum_{i=1}^{N} {\mu_i}^* y_i x_i = 0$
2. $\sum_{i=1}^{N} {\mu_i}^* y_i = 0$
3. $y_i ({w^*}^T x_i + {w_0}^*) \geq 1$
4. ${\mu_i}^* \geq 0$
5. ${\mu_i}^* \left(y_i ({w^*}^T x_i + {w_0}^*) - 1 \right) = 0$

Linearly separable case (5)

Substituting (1) and (2) into the Lagrange equation

$$L_d = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mu_i\mu_j y_i y_j {x_i}^T x_j + \sum_{i=1}^{N}\mu_i$$

Dual problem:

$$\text{Maximize } L_d$$
$$\text{s. t. } \sum_{i=1}^{N}\mu_i y_i = 0$$
$$\mu_i \geq 0$$

Linearly non-separable case (1)

If some of the samples are on the wrong side of the margin the feasible set of the problem will be empty.

We introduce slack variables ξ_i, which will allow some samples to be on the wrong side of the margin by ξ_i.

We also penalize over samples that are on the wrong side of the margin.

$$\text{minimize } \frac{\|w\|^2}{2} + C \sum_{i=1}^{N} \xi_i$$

$$\text{s. t. } y_i(w^T x_i + w_0) \geq 1 - \xi_i, \forall i$$

$$\xi_i \geq 0$$

Linearly separable case (2)

The Lagrange equation:

$$L_p = \frac{\|w\|^2}{2} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i(y_i(w^T x_i + w_0) - 1 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i$$

KKT conditions (differences from the linearly separable case highlighted in color):

$$1.\ \boldsymbol{w}^* - \sum_{i=1}^{N}\mu_i{}^* y_i \boldsymbol{x}_i = 0\,,\ 2.\ \sum_{i=1}^{N}\mu_i{}^* y_i = 0\,,$$

$$3.\ C = \mu_i{}^* + \beta_i{}^*,\ \ 4.\ y_i(\boldsymbol{w}^{*T}\boldsymbol{x}_i + w_0{}^*) \geq 1 - \xi_i,$$

$$5.\ \mu_i{}^* \geq 0, 6.\ \mu_i{}^*\left(y_i(\boldsymbol{w}^{*T}\boldsymbol{x}_i + w_0{}^*\right) - 1 + \xi_i) = 0,$$

$$7.\ \beta_i{}^* \geq 0, 8.\ \xi_i{}^* \geq 0, 9.\ \beta_i{}^*\,\xi_i{}^* = 0$$

Linearly non-separable case (3)

Substituting (1) and (2) into the Lagrange equation

$$L_d = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mu_i\mu_j y_i y_j {x_i}^T x_j + \sum_{i=1}^{N}\mu_i$$

Dual problem:

$$\text{Maximize } L_d$$
$$\text{s. t. } \sum_{i=1}^{N}\mu_i y_i = 0$$
$$C \geq \mu_i \geq 0$$

Support vectors

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \mu_i{}^* y_i \boldsymbol{x}_i$$

The above expression implies that x_i is being used in the derivation of $\boldsymbol{w}^*$ when the corresponding $\mu_i{}^* > 0$.

x_i with $\mu_i{}^* > 0$ are called support vectors. There are two types of them: bounded and unbounded.

Bounded support vectors are those with $\mu_i{}^* = C$. Thus, $\beta_i{}^* = 0$ (KKT condition 3) and $\xi_i > 0$

Unbounded support vectors are those with $\mu_i{}^* < C$. Thus, $\beta_i{}^* > 0$ (KKT condition 3) and $\xi_i = 0$

1-norm SVM

When the margin is computed using the 1-norm of w instead of the 2-norm used in traditional SVM this leads to a sparse form of w with only a subset of the vector coefficients being non-zero.

$$\text{minimize } \frac{|w|_1}{2} + C\sum_{i=1}^{N}\xi_i$$

$$\text{s. t. } y_i(w^T x_i + w_0) \geq 1 - \xi_i, \forall i$$

$$\xi_i \geq 0$$

1-norm SVM (2)

Solving this problem is equivalent to solving the following problem with auxiliary variables ν_i

$$\text{minimize } \frac{\sum_{i=1}^{N} \nu_i}{2} + C \sum_{i=1}^{N} \xi_i$$

$$\text{s. t. } y_i(w^T x_i + w_0) \geq 1 - \xi_i, \forall i$$

$$\xi_i \geq 0$$

$$-\nu_i \leq w_i \leq \nu_i, \forall i$$