Topic 5: SUPPORT VECTOR MACHINES

STAT 37710/CMSC 25400 Machine Learning

Risi Kondor, The University of Chicago

Regularized Risk Minimization (RRM)

Find the hypothesis $\widehat{f}$ by solving a problem of the form

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \left[\underbrace{\frac{1}{m} \sum_{i=1}^{m} \ell(f(x_i), y_i)}_{\text{training error}} + \underbrace{\lambda\, \Omega[f]}_{\text{regularizer}} \right]$$

- $\mathcal{F}$ can be quite a rich hypothesis space.
- The purpose of the regularizer is to avoid overfitting.
- λ is a tunable parameter.
- $\ell(\widehat{y}, y)$: loss function
- ℓ might or might not be the same loss as in $\mathcal{E}_{\text{true}}$.

[Tykhonov regularization] [Vapnik 1970's–]

Optimization: equality constraints

Problem:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x}) \qquad \text{subject to} \qquad g(\mathbf{x}) = c.$$

1. Form the **Lagrangian** $L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda\,(g(\mathbf{x}) - c)$.
2. The solution must be at a critical point of L. $\rightarrow$ Setting

$$\frac{\partial L(\mathbf{x}, \lambda)}{\partial x_i} = 0 \qquad i = 1, 2, \ldots, n.$$

   yields a curve of solutions $\mathbf{x} = \gamma(\lambda)$.
3. Reintroducing the constraint $g(\gamma(\lambda)) = c$ gives λ, hence the optimal $\mathbf{x}$.

Optimization: inequality constraints

Problem:

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}) \qquad \text{subject to} \qquad g(\mathbf{x}) \geq c.$$

1. Form the **Lagrangian** $L(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda\,(g(\mathbf{x}) - c)\,.$
2. Introduce the **dual function**

$$h(\lambda) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda).$$

3. Solve the dual problem

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, h(\lambda) \quad \text{subject to} \quad \lambda \geq 0.$$

4. The optimal $\mathbf{x}$ is $\inf_{\mathbf{x}} L(\mathbf{x}, \lambda^*)$ (assuming strong duality).

When f is a convex function and $g(\mathbf{x}) \geq c$ defines a convex region of space, this gives the global optimum.

Karush–Kuhn–Tucker conditions

At the optimal solution $\mathbf{x}^*$ of

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}) \qquad \text{subject to} \qquad g(\mathbf{x}) \geq c.$$

either

1. we are the boundary $\rightarrow$ $g(\mathbf{x}^*) = c$ or
2. we are at an interior point $\rightarrow$ $\lambda^* = 0$.

$\rightarrow$ **Complementary slackness**: $\lambda^* \left(g(\mathbf{x}^*) - c\right) = 0.$

Support Vector Machines

Linear classifiers

To apply RRM, go back to binary classification in $\mathbb{R}^n$ with a linear (affine) hyperplane:

Input space: $\mathcal{X} = \mathbb{R}^n$
Output space: $\mathcal{Y} = \{-1, +1\}$
Hypothesis:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b.$$

$$h(\mathbf{x}) = \operatorname{sgn}(f(\mathbf{x}))$$

(Note the sneaky difference between f and h)

Question: Of all possible hyperplanes that separate the data which one do we choose?

The margin

Recall, the **margin** of a point $(\mathbf{x}, y)$ to the hyperplane $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0$ (with $\|\mathbf{w}\| = 1$) is

$$y\,(\mathbf{w} \cdot \mathbf{x} + b).$$

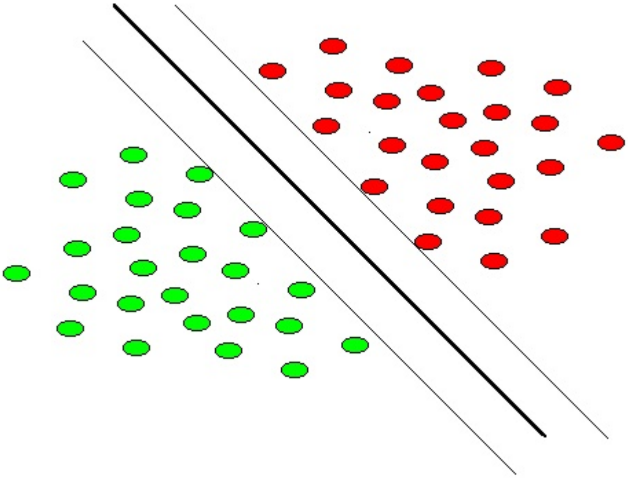
The margin of a dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ to f is

$$\min_i \; y_i\,(\mathbf{w} \cdot \mathbf{x}_i + b)\,.$$

In the case of the perceptron we saw that having a large margin is desirable.

IDEA: Choose $\mathbf{w}$ and b explicitly to maximize the margin! $\rightarrow$ **Support Vector Machines (SVM)**

Maximizing the margin



Choose the hyperplane that has the largest margin!

Hard Margin Support Vector Machine

Given a dataset $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$,

$$\underset{\|\mathbf{w}\|=1,\, b}{\text{maximize}} \;\; \delta \qquad \text{s.t.} \qquad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq \delta \quad \forall i.$$

Equivalent formulation: drop the $\|\mathbf{w}\| = 1$ constraint and solve

$$\underset{\mathbf{w},\, b}{\text{minimize}} \;\; \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{s.t.} \qquad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i.$$

The primal problem

The primal SVM optimization problem

$$\underset{\mathbf{w},b}{\text{minimize}} \; \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{s.t.} \qquad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \forall i$$

This is a nice convex optimization problem (a QP) with a unique minimum.
$\rightarrow$ Introduce a Lagrangian.

From primal to dual

$$\underset{\mathbf{w},b}{\text{minimize}} \ \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{s.t.} \qquad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad \forall i$$

Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \textstyle\sum_i \alpha_i(y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1)$$

$$\tfrac{\partial}{\partial w_i} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0 \qquad \Rightarrow \qquad \boxed{\mathbf{w} - \textstyle\sum_i \alpha_i y_i \mathbf{x}_i = 0}$$

$$\tfrac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0 \qquad \Rightarrow \qquad \textstyle\sum_i \alpha_i y_i = 0$$

Dual function:

$$L(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

The dual problem

The dual SVM optimization problem

$$\underset{\alpha_1,\ldots,\alpha_m}{\text{maximize}} \quad L(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

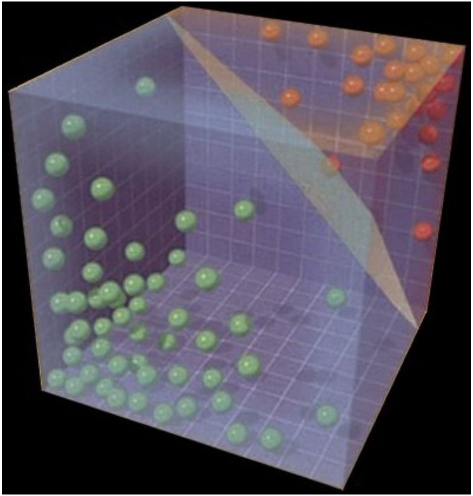
$$\text{subject to} \quad \sum_i y_i \alpha_i = 0 \quad \text{and} \quad \alpha_i \geq 0 \ \forall i$$

Still a QP, but in fewer variables, so easier to solve. In particular,

$$h(\mathbf{x}) = \operatorname{sgn}\Big[\sum_i \alpha_i y_i (\mathbf{x} \cdot \mathbf{x}_i) + b\Big] = \operatorname{sgn}\Big[\sum_i \gamma_i (\mathbf{x} \cdot \mathbf{x}_i) + b\Big],$$

where $\gamma_i = y_i \alpha_i$. $\rightarrow$ The solution lies in the span of the data, $\mathbf{w} = \sum_i \gamma_i \mathbf{x}_i$.

Support vector machine



Sparsity of support vectors

The KKT conditions prescribe that

$$\alpha_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1) = 0 \qquad \forall i$$

So $\alpha_i \neq 0$ only for those examples that lie exactly on the margin, and therefore only these "**support vectors**" influence the solution

$$h(\mathbf{x}) = \mathrm{sgn}\left[\sum_i \alpha_i y_i (\mathbf{x} \cdot \mathbf{x}_i) + b\right]$$

$\rightarrow$ Sparsity is a precious thing.

Question: But what about non-separable data? $\rightarrow$ Soft margin SVMs

The Soft Margin SVM

The primal SVM optimization problem

$$\underset{\mathbf{w},b,\xi_1,\ldots,\xi_m}{\text{minimize}} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_i \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \quad \forall i$$

The ξ_i's are called **slack variables** and C is a "softness parameter"

[Cortes & Vapnik, 1995]

From primal to dual

$$\underset{\mathbf{w},b,\xi_1,\dots,\xi_m}{\text{minimize}} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_i \xi_i \quad \text{s.t.} \quad y_i(\boldsymbol{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \quad \forall i$$

Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + \tfrac{C}{m}\sum_i \xi_i - \sum_i \alpha_i(y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 + \xi_i) - \sum_i \beta_i \xi_i$$

$$\tfrac{\partial}{\partial w_i} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0 \qquad \Rightarrow \qquad \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0$$

$$\tfrac{\partial}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0 \qquad \Rightarrow \qquad \sum_i \alpha_i y_i = 0$$

$$\tfrac{\partial}{\partial \xi_i} L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0 \qquad \Rightarrow \qquad \alpha_i + \beta_i = \tfrac{C}{m}$$

Soft margin SVM dual

The dual SVM optimization problem

$$\underset{\alpha_1,\ldots,\alpha_m}{\text{maximize}} \quad L(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{subject to} \quad \sum_i y_i \alpha_i = 0 \quad \text{and} \quad 0 \le \alpha_i \le \frac{C}{m} \ \forall i$$

SVM is just a form of RRM

At the optimum of the primal problem the slacks are as small as possible:

$$\xi_i = \max\{0, 1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + b)\} = \underbrace{(1 - y_i(\mathbf{w}\cdot\mathbf{x}_i + b))_{\geq 0}}_{\ell_{\text{hinge}}(\mathrm{w}\cdot\mathbf{x}_i, y_i)},$$

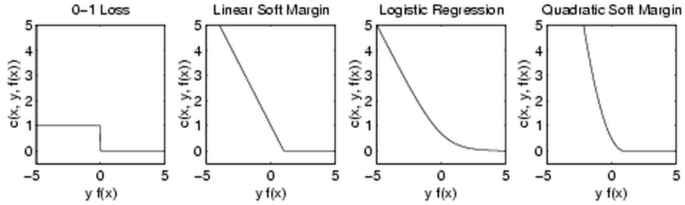
where $(z)_{\geq 0} = \max(0, z)$.

The soft-margin SVM finds

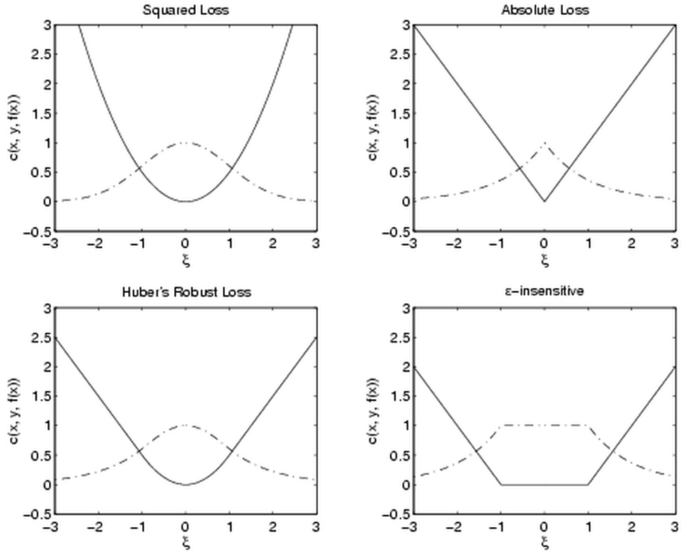
$$\widehat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \Bigg[\underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell_{\text{hinge}}(f(\mathbf{x}_i), y_i)}_{\text{empirical loss}} + \underbrace{\frac{1}{2C}\,\|\mathbf{w}\|^2}_{\text{regularizer}} \Bigg].$$

where $\mathcal{F}$ is the hypothesis space of $f(x) = \mathbf{w}\cdot\mathbf{x} + b$ linear functions.

Loss functions for classification



Loss functions for regression




SUPPORT VECTOR MACHINES
FREE VARIABLES!
END DUALITY GAP
Healthcare NOT Warfare