Machine learning: lecture 7

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Topics

- Support vector machines
  - separable case, formulation, margin
  - non-separable case, penalties, and logistic regression
  - dual solution, kernels
  - examples, properties

Support vector machine (SVM)

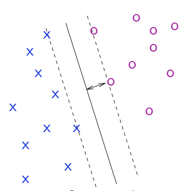
- When the training examples are *linearly separable* we can maximize a geometric notion of margin (distance to the boundary) by minimizing the regularization penalty

$$\|\mathbf{w}_1\|^2/2 = \sum_{i=1}^{d} w_i^2/2$$

subject to the classification constraints

$$y_i[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0$$

for $i = 1, \ldots, n$.

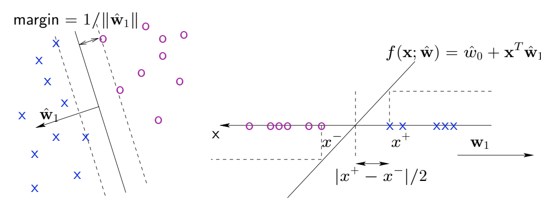


- The solution is defined only on the basis of a subset of examples or "support vectors"

SVM: separable case

- We minimize $\|\mathbf{w}_1\|^2/2 = \sum_{i=1}^{d} w_i^2/2$ subject to

$$y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \;\; i = 1, \ldots, n$$



- The resulting margin and the "slope" $\|\hat{\mathbf{w}}_1\|$ are inversely related

SVM: non-separable case

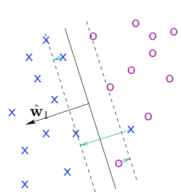
- When the examples are not linearly separable we can modify the optimization problem slightly to add a penalty for violating the classification constraints:
  We minimize

$$\|\mathbf{w}_1\|^2/2 + C\sum_{i=1}^{n} \xi_i$$

subject to relaxed classification constraints

$$y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 + \xi_i \geq 0,$$

for $i = 1, \ldots, n$. Here $\xi_i \geq 0$ are called "slack" variables.



SVM: non-separable case cont'd

- We can also write the SVM optimization problem more compactly as

$$C\sum_{i=1}^{n} \overbrace{\Big(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\Big)^+}^{\xi_i} + \|\mathbf{w}_1\|^2/2$$

where $(z)^+ = z$ if $z \geq 0$ and zero otherwise (i.e., returns the positive part).

SVM: non-separable case cont'd

- We can also write the SVM optimization problem more compactly as

$$C \sum_{i=1}^{n} \overbrace{\left(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right)^+}^{\xi_i} + \|\mathbf{w}_1\|^2/2$$

where $(z)^+ = z$ if $z \geq 0$ and zero otherwise (i.e., returns the positive part).

- This is equivalent to regularized empirical loss minimization

$$\frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right)^+ + \lambda \|\mathbf{w}_1\|^2/2$$

where $\lambda = 1/nC$ is the regularization parameter.

SVM vs logistic regression

- When viewed from the point of view of regularized empirical loss minimization, SVM and logistic regression appear quite similar:

$$\text{SVM:} \quad \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right)^+ + \lambda \|\mathbf{w}_1\|^2/2$$

$$\text{Logistic:} \quad \frac{1}{n} \sum_{i=1}^{n} \overbrace{-\log g\left(y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right)}^{-\log P(y_i|\mathbf{x},\mathbf{w})} + \lambda \|\mathbf{w}_1\|^2/2$$

where $g(z) = (1 + \exp(-z))^{-1}$ is the logistic function.

(Note that we have transformed the problem maximizing the penalized log-likelihood into minimizing negative penalized log-likelihood.)

SVM vs logistic regression cont'd

- The difference comes from how we penalize "errors":

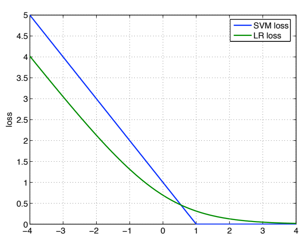
$$\text{Both:} \quad \frac{1}{n} \sum_{i=1}^{n} \text{Loss}\Big(\overbrace{y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]}^{z}\Big) + \lambda \|\mathbf{w}_1\|^2/2$$

- SVM:

  $\text{Loss}(z) = (1 - z)^+$

- Regularized logistic reg:

  $\text{Loss}(z) = \log(1 + \exp(-z))$



SVM: solution, Lagrange multipliers

- Back to the separable case: how do we solve

$$\min\ \|\mathbf{w}_1\|^2/2 \quad \text{subject to}$$
$$y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \quad i = 1, \ldots, n$$

SVM: solution, Lagrange multipliers

- Back to the separable case: how do we solve

$$\min\ \|\mathbf{w}_1\|^2/2 \quad \text{subject to}$$
$$y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \quad i = 1, \ldots, n$$

- Let start by representing the constraints as losses

$$\max_{\alpha \geq 0} \alpha\left(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right) = \begin{cases} 0, & y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0 \\ \infty, & \text{otherwise} \end{cases}$$

SVM: solution, Lagrange multipliers

- Back to the separable case: how do we solve

$$\min\ \|\mathbf{w}_1\|^2/2 \quad \text{subject to}$$
$$y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0, \quad i = 1, \ldots, n$$

- Let start by representing the constraints as losses

$$\max_{\alpha \geq 0} \alpha\left(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right) = \begin{cases} 0, & y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1] - 1 \geq 0 \\ \infty, & \text{otherwise} \end{cases}$$

and rewrite the minimization problem in terms of these

$$\min_{\mathbf{w}} \left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \max_{\alpha_i \geq 0} \alpha_i\left(1 - y_i\,[w_0 + \mathbf{x}_i^T \mathbf{w}_1]\right) \right\}$$

SVM: solution, Lagrange multipliers

- Back to the separable case: how do we solve

$$\min \|\mathbf{w}_1\|^2/2 \;\; \text{subject to}$$
$$y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1] - 1 \geq 0, \;\; i = 1, \ldots, n$$

- Let start by representing the constraints as losses

$$\max_{\alpha \geq 0} \alpha\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) = \begin{cases} 0, & y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1] - 1 \geq 0 \\ \infty, & \text{otherwise} \end{cases}$$

and rewrite the minimization problem in terms of these

$$\min_{\mathbf{w}} \left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \max_{\alpha_i \geq 0} \alpha_i\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) \right\}$$
$$= \min_{\mathbf{w}} \max_{\{\alpha_i \geq 0\}} \left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \alpha_i\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) \right\}$$

SVM solution cont'd

- We can then swap 'max' and 'min':

$$\min_{\mathbf{w}} \max_{\{\alpha_i \geq 0\}} \left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \alpha_i\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) \right\}$$
$$\stackrel{?}{=} \max_{\{\alpha_i \geq 0\}} \min_{\mathbf{w}} \underbrace{\left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \alpha_i\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) \right\}}_{J(\mathbf{w};\alpha)}$$

As a result we have to be able to minimize $J(\mathbf{w};\alpha)$ with respect to parameters $\mathbf{w}$ for any fixed setting of the Lagrange multipliers $\alpha_i \geq 0$.

SVM solution cont'd

- We can then swap 'max' and 'min':

$$\min_{\mathbf{w}} \max_{\{\alpha_i \geq 0\}} \left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \alpha_i\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) \right\}$$
$$\stackrel{?}{=} \max_{\{\alpha_i \geq 0\}} \min_{\mathbf{w}} \underbrace{\left\{ \|\mathbf{w}_1\|^2/2 + \sum_{i=1}^{n} \alpha_i\big(1 - y_i\,[w_0 + \mathbf{x}_i^T\mathbf{w}_1]\big) \right\}}_{J(\mathbf{w};\alpha)}$$

We can find the optimal $\hat{\mathbf{w}}$ as a function of $\{\alpha_i\}$ by setting the derivatives to zero:

$$\frac{\partial}{\partial \mathbf{w}_1} J(\mathbf{w};\alpha) = \mathbf{w}_1 - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$
$$\frac{\partial}{\partial w_0} J(\mathbf{w};\alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

SVM solution cont'd

- We can then substitute the solution

$$\frac{\partial}{\partial \mathbf{w}_1} J(\mathbf{w};\alpha) = \mathbf{w}_1 - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$
$$\frac{\partial}{\partial w_0} J(\mathbf{w};\alpha) = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

back into the objective and get (after some algebra):

$$\max_{\substack{\alpha_i \geq 0 \\ \sum_i \alpha_i y_i = 0}} \left\{ \|\hat{\mathbf{w}}_1\|^2/2 + \sum_{i=1}^{n} \alpha_i\big(1 - y_i\,[\hat{w}_0 + \mathbf{x}_i^T\hat{\mathbf{w}}_1]\big) \right\}$$
$$= \max_{\substack{\alpha_i \geq 0 \\ \sum_i \alpha_i y_i = 0}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i^T \mathbf{x}_j) \right\}$$

SVM solution: summary

- We can find the optimal setting of the Lagrange multipliers α_i by maximizing

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i^T \mathbf{x}_j)$$

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$. Only α_i's corresponding to "support vectors" will be non-zero.

SVM solution: summary

- We can find the optimal setting of the Lagrange multipliers α_i by maximizing

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i^T \mathbf{x}_j)$$

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$. Only α_i's corresponding to "support vectors" will be non-zero.

- We can make predictions on any new example $\mathbf{x}$ according to the sign of the discriminant function

$$\hat{w}_0 + \mathbf{x}^T\hat{\mathbf{w}}_1$$

SVM solution: summary

- We can find the optimal setting of the Lagrange multipliers α_i by maximizing

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i^T \mathbf{x}_j)$$

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$. Only α_i's corresponding to "support vectors" will be non-zero.

- We can make predictions on any new example $\mathbf{x}$ according to the sign of the discriminant function

$$\hat{w}_0 + \mathbf{x}^T \hat{\mathbf{w}}_1 = \hat{w}_0 + \mathbf{x}^T \big(\sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i \big)$$

SVM solution: summary

- We can find the optimal setting of the Lagrange multipliers α_i by maximizing

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (\mathbf{x}_i^T \mathbf{x}_j)$$

subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$. Only α_i's corresponding to "support vectors" will be non-zero.

- We can make predictions on any new example $\mathbf{x}$ according to the sign of the discriminant function

$$\hat{w}_0 + \mathbf{x}^T \hat{\mathbf{w}}_1 = \hat{w}_0 + \mathbf{x}^T \big(\sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i \big) = \hat{w}_0 + \sum_{i \in SV} \hat{\alpha}_i y_i (\mathbf{x}^T \mathbf{x}_i)$$

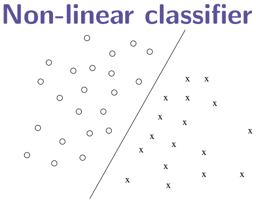
Non-linear classifier

- So far our classifier can make only linear separations
- As with linear regression and logistic regression models, we can easily obtain a non-linear classifier by first mapping our examples $\mathbf{x} = [x_1 \; x_2]$ into longer feature vectors $\phi(\mathbf{x})$

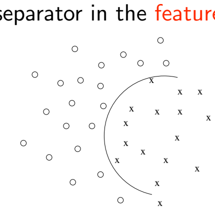
$$\phi(\mathbf{x}) = [x_1^2 \;\; x_2^2 \;\; \sqrt{2}x_1 x_2 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2 \;\; 1]$$

and then applying the linear classifier to the new feature vectors $\phi(\mathbf{x})$

Non-linear classifier



Linear separator in the feature ϕ-space



Non-linear separator in the original $\mathbf{x}$-space

Feature mapping and kernels

- Let's look at the previous example in a bit more detail

$$\mathbf{x} \rightarrow \phi(\mathbf{x}) = [x_1^2 \;\; x_2^2 \;\; \sqrt{2}x_1 x_2 \;\; \sqrt{2}x_1 \;\; \sqrt{2}x_2 \;\; 1]$$

- The SVM classifier deals only with inner products of examples (or feature vectors). In this example,

$$\begin{aligned} \phi(\mathbf{x})^T \phi(\mathbf{x}') &= x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1 x_2 x_1' x_2' + 2x_1 x_1' + 2x_2 x_2' + 1 \\ &= (1 + x_1 x_1' + x_2 x_2')^2 \\ &= \big(1 + (\mathbf{x}^T \mathbf{x}')\big)^2 \end{aligned}$$

so the inner products can be evaluated without ever explicitly constructing the feature vectors $\phi(\mathbf{x})$!

- $K(\mathbf{x}, \mathbf{x}') = \big(1 + (\mathbf{x}^T \mathbf{x}')\big)^2$ is a *kernel function* (inner product in the feature space)

Examples of kernel functions

- **Linear kernel**

$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')$$

- **Polynomial kernel**

$$K(\mathbf{x}, \mathbf{x}') = \big(1 + (\mathbf{x}^T \mathbf{x}')\big)^p$$

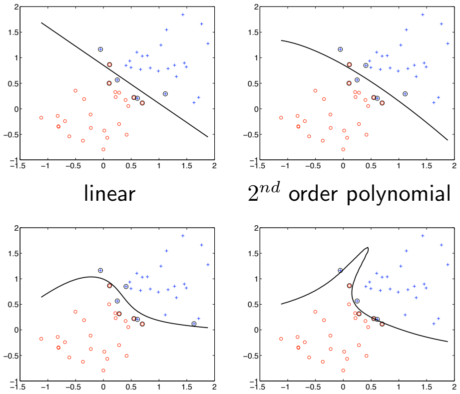
where $p = 2, 3, \ldots$. To get the feature vectors we concatenate all up to p^{th} order polynomial terms of the components of $\mathbf{x}$ (weighted appropriately)

- **Radial basis kernel**

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \right)$$

In this case the feature space is infinite dimensional function space (use of the kernel results in a *non-parametric* classifier).

SVM examples



linear | 2^{nd} order polynomial | 4^{th} order polynomial | 8^{th} order polynomial