Support Vector Machines II

CSci 5525: Machine Learning

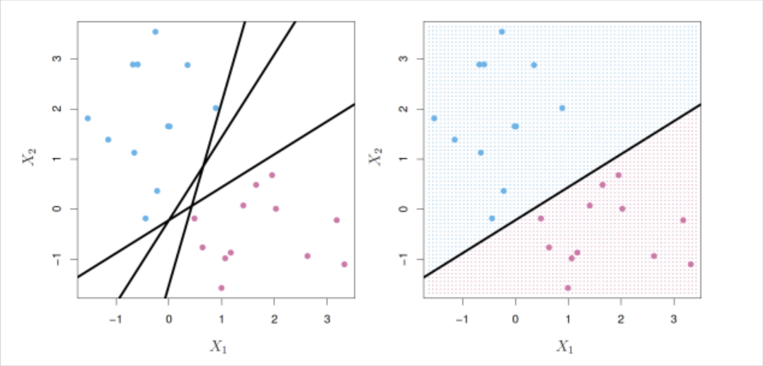
Instructor: Nicholas Johnson

September 29, 2020

Announcements

- HW1 due Thu Oct 1
- HW2 will be posted Thu Oct 1 (due Oct 15)

Linear Classification

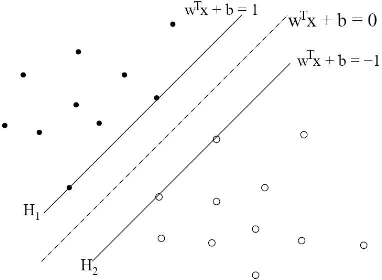


Max Margin

- Max margin idea: select predictor that maximizes distance between data points and decision boundary
- Linear predictor: $\mathbf{w}^\top \mathbf{x} + b$
- Decision boundary: $\{\mathbf{x} \in \mathbb{R}^p : \mathbf{w}^\top \mathbf{x} + b = 0\}$ (hyperplane)
- When perfectly classified we have

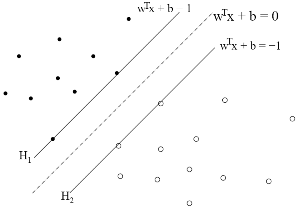
$$(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, 1\} : y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 0 \; \forall i$$

Max Margin



- Distance of $\mathbf{x}_i$ to decision boundary $= \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|}$
- Smallest distance to decision boundary: $\min_i \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|}$
- Main idea: *Choose* $\mathbf{w}$ *to maximize class separation*

Linear SVM: Separable Case



$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{such that} \quad \forall i,\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

- The choice of “1” as a constant is *wlog*
- The main problem is a “quadratic program”
- Computes linear classifier with largest margin - the support vector machine (SVM) classifier
- Solution is unique (why?)

Linear SVM: Non-Separable Case

- Separability assumption: $\exists \mathbf{w}, \forall i \; y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$
- If not true, the problem formulation is infeasible
- For the general case, we will introduce *slack variables*

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall i$$

- In general, the problem can be formulated as

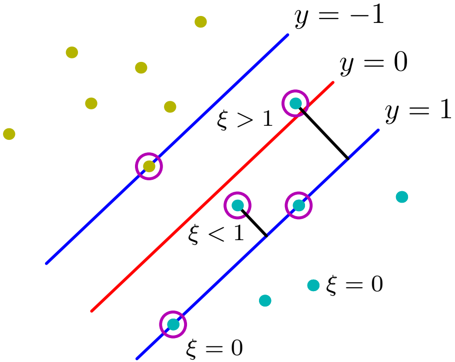
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_i \xi_i \quad \text{such that}$$

$$\forall i, \; y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$$

$$\forall i, \; \xi_i \geq 0$$

- Perspective: constrained optimization

Linear SVM: Non-Separable Case



Constrained Optimization

- The inequality constrained optimization problem

$$\begin{aligned} \text{minimize}_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0 \qquad i = 1, \ldots, m \end{aligned}$$

- For each constraint, introduce Lagrangian multiplier $\lambda_j \geq 0$
- The Lagrangian

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}) &= f(\mathbf{x}) + \boldsymbol{\lambda}^\top h(\mathbf{x}) \\ &= f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) \end{aligned}$$

- Consider the problem $\max_{\boldsymbol{\lambda}} \max_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})$
  - let $\mathbf{x}^* = \text{argmin}_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda})$ and
  - $\boldsymbol{\lambda}^* = \text{argmax}_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})$
- Consider the following derivation:

$$\begin{aligned}
\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) &= \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}^*) \\
&\leq L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\
&\leq \max_{\boldsymbol{\lambda}} L(\mathbf{x}^*, \boldsymbol{\lambda}) \\
&= \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda})
\end{aligned}$$

- The relationship $\max \min \leq \min \max$ is called weak duality

Constrained Optimization

- Under mild conditions such as Slater's condition (e.g., in quadratic programs) we have strong duality

$$\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda})$$

- Under strong duality we have

$$\begin{aligned}
f(\mathbf{x}^*) &= \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) && \text{(definition of } \mathbf{x}^*\text{)} \\
&= \max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) && \text{(strong duality)} \\
&= \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}^*) && \text{(definition of } \boldsymbol{\lambda}^*\text{)} \\
&\leq L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \\
&= f(\mathbf{x}^*) + \sum_i \lambda_i^* h_i(\mathbf{x}^*)
\end{aligned}$$

Complementary Slackness

- Since $\mathbf{x}^*$ is feasible then $\lambda_i^* h_i(\mathbf{x}^*) = 0 \; \forall i$
- This implies the last inequality must hold with equality
  - $\lambda_i^* > 0 \implies h_i(\mathbf{x}^*) = 0$
  - $h_i(\mathbf{x}^*) < 0 \implies \lambda_i^* = 0$

Stationarity

- $\mathbf{x}^*$ is minimizer of $L(\mathbf{x}, \boldsymbol{\lambda}^*)$ therefore it has gradient zero

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \nabla f(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla h_i(\mathbf{x}^*) = 0$$

Feasibility

- Primal feasibility: $h_i(\mathbf{x}^*) \leq 0 \; \forall i$
- Dual feasibility: $\lambda_i \geq 0 \; \forall i$

The conditions are sufficient for a convex problem

Dual Formulation

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i \quad \text{such that}$$
$$\forall i,\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$$
$$\forall i,\ \xi_i \geq 0$$

- Rewrite each constraint as $1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0$ with dual variable $\lambda_i \geq 0$
- For each constraint $\xi_i \geq 0$ we introduce dual variable $\alpha_i \geq 0$
- Variables $\mathbf{w}$ and ξ are called primal variables
- Langrangian is:

$$L(\mathbf{w}, \xi, \lambda, \alpha) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \lambda_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \alpha_i \xi_i$$

- Now we can apply KKT conditions to characterize the SVM solution

Dual Formulation

$$L(\mathbf{w}, \xi, \lambda, \alpha) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i + \sum_i \lambda_i(1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_i \alpha_i \xi_i$$

- Applying stationarity condition $\nabla_{\mathbf{w},\xi} L(\mathbf{w}^*, \xi^*, \lambda^*, \alpha^*) = 0$ we get

$$\mathbf{w} = \sum_i y_i \lambda_i^* \mathbf{x}_i$$

$$C - \lambda_i^* - \alpha_i^* = 0 \ \forall i$$

Dual Formulation

- Plugging these into L we get

$$L(\mathbf{w}, \xi, \lambda, \alpha) = \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

- Optimization problem becomes

$$\max_{\alpha,\lambda} \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \quad \text{such that}$$
$$C = \lambda_i + \alpha_i, \ \forall i$$
$$\lambda_i, \alpha_i \geq 0, \ \forall i$$

- Quadratic program: quadratic objective functions with linear constraints

Dual Formulation

- With optimal solution λ^* we know from the KKT conditions:

$$\mathbf{w}^* = \sum_i y_i \lambda_i^* \mathbf{x}_i = \sum_{i:\lambda_i^* > 0} y_i \lambda_i^* \mathbf{x}_i$$

- Any point i with $\lambda_i^* > 0$ is called a support vector, hence the name support vector machine

Dual Formulation

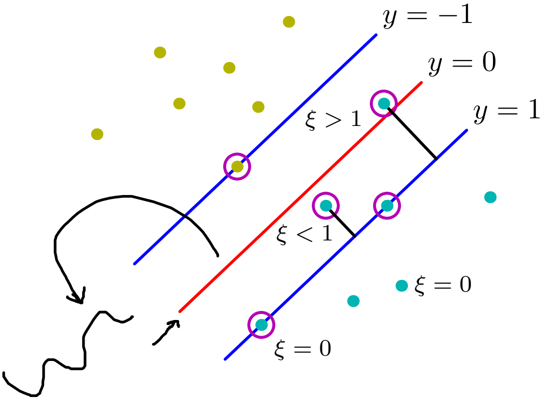
- Applying complementary slackness condition we get

$$\lambda_i^*(1 - \xi_i^* - y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b)) = 0, \ \forall i$$
$$\alpha_i^* \xi_i^* = 0, \ \forall i$$

- For any support vector we have $1 - \xi_i^* = y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b)$
- b can similarly be obtained using KKT conditions

- If $\xi_i^* = 0$ then $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b) = 1$
  - Point i is $1/\|\mathbf{w}\|$ away from decision boundary
- If $\xi_i^* < 0$ then $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b) \in (0, 1)$
  - Point i is correctly classified but close to decision boundary
- If $\xi_i^* > 0$ then $y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b) < 0$
  - Point i is incorrectly classified

Linear SVM: Non-Separable Case



SVM Prediction

- For any future point $\mathbf{x}$ the prediction is

$$\operatorname{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle + b) = \operatorname{sign}\left(\overbrace{\sum_{i:\lambda_i > 0} y_i \lambda_i^* \mathbf{x}_i^\top}^{= w^*} \mathbf{x} + b \right)$$

- Note: prediction in terms of dot products $\mathbf{x}_i^\top \mathbf{x}$, dual also in terms of dot products $\mathbf{x}_i^\top \mathbf{x}_j$

Non-linear SVMs

- All important equations have dot-products
  - Dual is expressed in terms of $\mathbf{x}_i^\top \mathbf{x}_j$
  - The predictions are in terms of $\mathbf{x}_i^\top \mathbf{x}$
- How to get a non-linear classifier:
  - Map $\mathbf{x}$ to some (higher dimensional) space $\Phi : \mathbb{R}^p \mapsto \mathcal{H}$ feature space
  - The derived feature vectors are $\Phi(\mathbf{x}_i), \forall i$
  - The dot products are $\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_j)$
- Kernel function allows implicit calculation of dot-products

$K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_j)$$

- Learn a linear max margin separator in $\mathcal{H}$
- The final prediction function

$$\mathrm{sign}\left(f(\mathbf{x})\right) = \Big[\underbrace{\sum_{i:\lambda_i^*>0} y_i \lambda_i^* \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}) + b}_{f(x)}\Big] = \Big[\sum_{i:\lambda_i^*>0} y_i \lambda_i^* K(\mathbf{x}_i, \mathbf{x}) + b\Big]$$

Non-linear SVMs

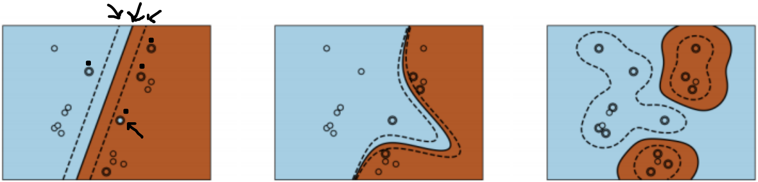


Figure : From left to right: decision boundaries of kernel SVM with linear, 3rd degree polynomial, and RBF kernels.

SRM View of SVM

- We can also view SVM through the SRM lens
- Class of predictors $\mathcal{C}$ is the set of linear/affine functions $\mathbf{w}^\top \mathbf{x} + b$
- Loss function: hinge loss $\max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$ 1.5
- Regularizer: L_2 norm squared $\|\mathbf{w}\|_2^2$ <0 $-\frac{1}{2}$
- SRM optimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

$\min \quad \ell(f(x)) + \lambda R(\cdot)$

Multiclass SVM

- SVM is inherently used for binary classification
- Two approaches for multiclass SVM: *one-against-all* and *one-against-one*

Multiclass SVM: One-against-all

- Solve k binary classification problems
- Classify class j against all other classes
- Given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, construct k datasets $\mathcal{D}_1, \ldots, \mathcal{D}_k$ where $\mathcal{D}_j = \{(\mathbf{x}_i, \mathbb{1}(y_i = j))\}_{i=1}^n$
- Run SVM k times on each dataset to obtain $\mathbf{w}_j$ and b_j
- For sample $\mathbf{x}$ predict label as $\hat{y} = \text{argmax}_j \, \mathbf{w}_j^\top \mathbf{x} + b_j$

Multiclass SVM: One-against-~~all~~ one

- Run SVM $k(k-1)/2$ times for every pair of labels j, j'
- Learn $\mathbf{w}_{jj'}$ that classifies the two classes using subset of data with labels j, j'
- For each sample $\mathbf{x}$, $\mathbf{w}_{jj'}$ "votes" for either label j or j'
- Predict class with highest votes given by $\mathbf{w}_{jj'}$

Multiclass SVM

- Another idea similar to one-against-all is to train $\mathbf{w}_1, \ldots, \mathbf{w}_k$ simultaneously
- Multiclass SVM optimization problem is:

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_k} \frac{1}{2}\sum_{j=1}^{k} \|\mathbf{w}_j\|_2^2 + C\sum_{i=1}^{n} \xi_i \quad \text{such that}$$

$$\mathbf{w}_{y_i}^\top \mathbf{x}_i \geq \mathbf{w}_j^\top \mathbf{x}_i + 1 - \xi_i, \ \forall i, \forall j \neq y_i$$

$$\xi_i \geq 0, \ \forall i$$