Support Vector Machines, Constrained Optimization

CSci 5525: Machine Learning

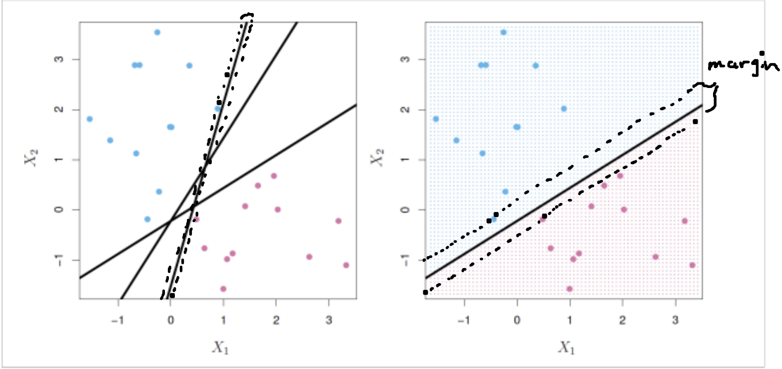
Instructor: Nicholas Johnson

September 24, 2020

Announcements

- HW1 posted (due Thu Oct 1 – 9 days)
- Project proposals due tonight at 11:59 PM CDT
  - No more than 1 page (problem to solve, initial solution idea, data/tools needed, etc.)
  - Submit via Canvas

Linear Classification

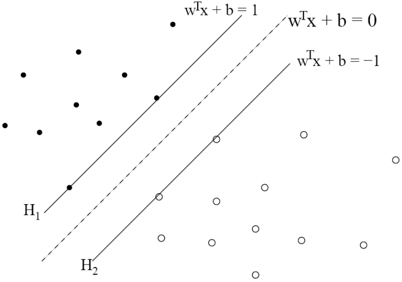


Max Margin

- 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*

Max Margin

- Main idea can be formulated as

$$\max_{\mathbf{w}} \min_i \frac{y_i(\mathbf{w}^\top \mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

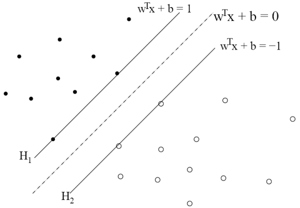
- Rescaling does not change optimal $\mathbf{w}$
- Suffices to consider $\mathbf{w}$ such that $\min_i y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$ to get

$$\max_{\mathbf{w}} \frac{1}{\|\mathbf{w}\|_2} \quad \text{such that} \quad \min_i y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$$

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

- Why are these two problems equivalent?

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

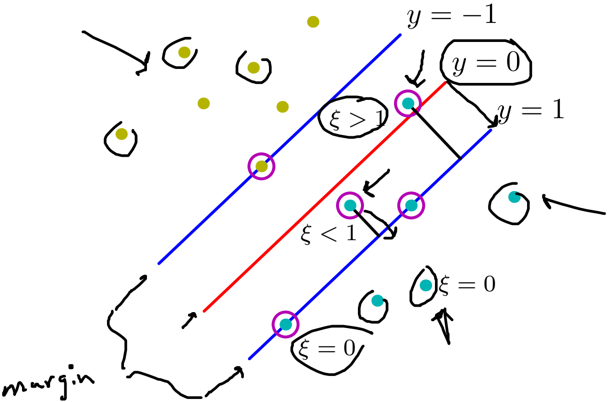
- Separability assumption: $\exists \mathbf{w}, \forall i \; y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$
- If not true, the problem formulation is infeasible
- For the general case, we will introduce *slack variables*

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

- In general, the problem can be formulated as

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

Linear SVM: Non-Separable Case



Linear SVM: Non-Separable Case

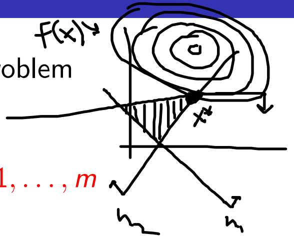
$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_i \xi_i \quad \text{such that}$$

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

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

- Note that $\sum_i \xi_i$ is an upper bound on the training error
- $\xi_i/\|\mathbf{w}\|_2$ is distance sample i needs to move to satisfy $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$
- Perspective: constrained optimization

Constrained Optimization



- The inequality constrained optimization problem

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

  - Domain $\mathcal{D} = \mathrm{dom}(f) \cap \bigcap_{i=1}^{m} \mathrm{dom}(h_i)$
  - Called the "primal" or primal problem
- Feasible set $\mathcal{F} \subseteq \mathcal{D}$: $\mathbf{x} \in \mathcal{F}$ satisfies $h_i(\mathbf{x}) \le 0$
- For each constraint, introduce Lagrangian multiplier $\lambda_i \ge 0$
- The Lagrangian

$h_i(x) > 0$

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

Constrained Optimization

why?

- $\max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda}) = \infty$ whenever $\mathbf{x}$ violates one of the constraints
- Therefore, solution to $\min_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda})$ is same as solution to constrained problem (why?)
- Consider the problem $\max_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})$
  - let $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \max_{\boldsymbol{\lambda}} L(\mathbf{x}, \boldsymbol{\lambda})$ and
  - $\boldsymbol{\lambda}^* = \operatorname{argmax}_{\boldsymbol{\lambda}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})$
- 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

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

$=0$

Constrained Optimization

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: $\lambda_i \geq 0 \ \forall i$

Necessary conditions satisfied by any primal and dual optimal pairs $\tilde{\mathbf{x}}$ and $(\tilde{\lambda}, \cancel{\tilde{x}})$

- Primal Feasibility:

$$h_i(\tilde{\mathbf{x}}) \leq 0, \ \forall i$$

- Dual Feasibility:

$$\tilde{\lambda}_i \geq 0, \forall i$$

- Complementary Slackness:

$$\tilde{\lambda}_i h_i(\tilde{\mathbf{x}}) = 0, \ \forall i$$

- Stationarity:

$$\nabla f(\tilde{\mathbf{x}}) + \sum_i \tilde{\lambda}_i \nabla h_i(\tilde{\mathbf{x}}) = 0$$

- The conditions are sufficient for a convex problem