Support Vector Machines

Gautam Kamath

Hard-Margin SVM

Setting

- Given (x_i, y_i) pairs, $x_i \in \mathbf{R}^d$, $y_i \in \{-1, +1\}$
- Dataset is linearly separable
    - (for now, while we're considering *hard-margin SVMs)*
- (Draw perceptron, highlight non-uniqueness drawback)
- Support Vector Machines (SVMs) try to find the "best" solution

Margins

- Suppose we have a perceptron solution w', b'
- If $\|w'\|_2 = 1$, then distance from point i to hyperplane is
$$\gamma_i = y_i(\langle w', x_i \rangle + b')$$
- (Draw picture)
- Recall margin of a dataset (wrt solution w', b') is $\gamma = \min_i \gamma_i$
- Perceptron tries to find any w', b' such that $\min_i \gamma_i \geq 0$
- SVM: Like perceptron, but try to maximize margin

Deriving SVM problem

- SVM: Like perceptron, but try to maximize margin

$$\max_{w',b'} \gamma,\ \text{s.t.}\ \|w'\|_2 = 1, y_i(\langle w', x_i\rangle + b') \geq \gamma \text{ for all } i$$

- Substitute $w' = \gamma w$, $b' = \gamma b$

$$\max_{\gamma w,\gamma b} \gamma,\ \text{s.t.}\ \|w\|_2 = 1/\gamma, y_i(\langle \gamma w, x_i\rangle + \gamma b) \geq \gamma \text{ for all } i$$

$$\max_{\gamma w,\gamma b} \gamma,\ \text{s.t.}\ \|w\|_2 = 1/\gamma, y_i(\langle w, x_i\rangle + b) \geq 1 \text{ for all } i$$

$$\max_{\gamma w,\gamma b} \frac{1}{\|w\|_2},\ \text{s.t.}\ y_i(\langle w, x_i\rangle + b) \geq 1 \text{ for all } i$$

$$\min_{w,b} \frac{1}{2}\|w\|_2^2\ \text{s.t.}\ y_i(\langle w, x_i\rangle + b) \geq 1 \text{ for all } i$$

Hard-Margin SVM problem

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 \ \text{ s.t. } y_i(\langle w, x_i \rangle + b) \geq 1 \text{ for all } i$$

- Instead of keeping $\|w\|_2$ fixed and maximizing γ, do opposite
- Let $\widehat{y_i} = \langle w, x_i \rangle + b$. Note that sign of $\widehat{y_i}$ gives prediction

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 \ \text{ s.t. } y_i \widehat{y_i} \geq 1 \text{ for all } i$$

- Compare with (weird writing of) perceptron's objective

$$\min_{w,b} 0 \ \text{ s.t. } y_i \widehat{y_i} \geq 1 \text{ for all } i$$

- Recall regularization
- Optimization?

Deriving dual formulation of SVM

Primal formulation of SVM:

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 \ \text{s.t.}\, y_i(\langle w, x_i\rangle + b) \geq 1 \text{ for all } i$$

Convert constrained optimization to unconstrained optimization

$$\max_{\alpha \in \mathbf{R}^n, \alpha \geq 0} \min_{w,b} \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i(y_i(\langle w, x_i\rangle + b) - 1)$$

Lagrange multiplier – adds penalty to objective for each constraint

Deriving dual formulation of SVM

$$\max_{\alpha \in \mathbf{R}^n, \alpha \geq 0} \min_{w,b} \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i (\langle w, x_i \rangle + b) - 1)$$

Fix some α for now, solve inner minimization. How? Set gradient $= 0$!

$$\frac{\partial}{\partial b} = -\sum_i \alpha_i y_i = 0, \qquad \frac{\partial}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0$$

Substitute into above and rearrange…

$$\min_{\alpha \in \mathbf{R}^n, \alpha \geq 0} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_i \alpha_i \;\; \text{s.t.} \sum_i \alpha_i y_i = 0$$

Dual formulation of SVM. Only depends on x_i's via dot products!!

Interpreting SVM solutions

$$\max_{\alpha \in \mathbf{R}^n, \alpha \geq 0} \min_{w,b} \frac{1}{2} \|w\|_2^2 - \sum_{i=1}^{n} \alpha_i (y_i(\langle w, x_i \rangle + b) - 1)$$

Property known as "complementary slackness" implies

$$\alpha_i (y_i(\langle w, x_i \rangle + b) - 1) = 0 \text{ for all } i \in [n]$$

Thus, either

$$a_i = 0 \quad \text{or} \quad (y_i(\langle w, x_i \rangle + b) = 1$$

If $a_i > 0$, these points are called *support vectors* (draw picture)

Note solution w can be written as combination of support vectors

$$w = \sum \alpha_i y_i x_i$$

Soft-Margin SVM

From Hard to Soft Margin SVMs

- (Draw non-separable, draw one outlier)
- Hard-margin:

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 \ \text{s.t.}\, y_i(\langle w, x_i \rangle + b) \geq 1 \text{ for all } i$$

- Equivalently, if we let $\widehat{y_i} = \langle w, x_i \rangle + b$

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 \ \text{s.t.}\ 1 - y_i \widehat{y_i} \leq 0 \text{ for all } i$$

- Soft-margin

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 + C \sum_i \max(0, 1 - y_i \widehat{y_i})$$

Interpreting Soft Margin

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 + C \sum_i \max(0, 1 - y_i \widehat{y_i})$$

- (draw cases of $y_i \widehat{y_i}$, versus hard-margin case)
  - If $1 - y_i \widehat{y_i} \leq 0$, then on the correct side of margin
  - If $0 \leq y_i \widehat{y_i} \leq 1$, correctly classified but within margin
  - If $y_i \widehat{y_i} \leq 0$, incorrectly classified
- (draw hinge loss versus 0-1 loss, perceptron)
- If $C = 0$, ignore data, if $C = \infty$, hard-margin SVM

Deriving dual formulation of SVM

$$\min_{w,b} \frac{1}{2} \|w\|_2^2 + C \sum_i \max(0, 1 - y_i \hat{y_i})$$

- Define "slack variables" γ_i
  - Interpretation: how far on wrong side of margin is point? (draw hinge loss)

$$\min_{w,b,\gamma} \frac{1}{2} \|w\|_2^2 + C \sum_i \gamma_i \ \text{s.t.} \max(0, 1 - y_i \hat{y_i}) \leq \gamma_i \text{ for all } i$$

- Break into two parts

$$\min_{w,b,\gamma} \frac{1}{2} \|w\|_2^2 + C \sum_i \gamma_i \ \text{s.t.}\ 0 \leq \gamma_i \text{ and } 1 - y_i \hat{y_i} \leq \gamma_i \text{ for all } i$$

Deriving dual formulation of SVM

$$\min_{w,b,\gamma} \frac{1}{2}\|w\|_2^2 + C\sum_i \gamma_i \text{ s.t. } 0 \le \gamma_i \text{ and } 1 - y_i\hat{y}_i \le \gamma_i \text{ for all } i$$

- Introduce dual variables and take Lagrangian

$$\max_{\alpha,\beta\in\mathbf{R}^n,\alpha,\beta\ge 0} \min_{w,b,\gamma} \frac{1}{2}\|w\|_2^2 + \sum_i \left(C\gamma_i + \alpha_i(1 - y_i\hat{y}_i - \gamma_i) - \beta_i\gamma_i\right)$$

- Take derivative of inner problem, set to 0, substitute, simplify… (exercise)

$$\min_{\alpha\in\mathbf{R}^n, C\ge\alpha\ge 0} \sum_i\sum_j \alpha_i\alpha_j y_i y_j\langle x_i, x_j\rangle - \sum_i \alpha_i \text{ s.t. } \sum_i \alpha_i y_i = 0$$

- Dual formulation of SVM. Only depends on x_i's via dot products!!

Interpreting SVM solutions

$$\min_{\alpha \in \mathbf{R}^n, C \geq \alpha \geq 0} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle - \sum_i \alpha_i \ \text{s.t.} \sum_i \alpha_i y_i = 0$$

- "complementary slackness" implies (after substitution $\beta_i = C - \alpha_i$)
  $$\alpha_i (1 - y_i \widehat{y_i} - \gamma_i) = 0 \text{ and } (C - \alpha_i)\gamma_i = 0$$
- Suppose $\alpha_i = 0$. Then $\gamma_i = 0$, point is on right side of margin (draw)
- Suppose $\alpha_i > 0$. Then $1 - y_i \widehat{y_i} - \gamma_i = 0$.
  - If $\gamma_i = 0$ (i.e., point sitting on margin), then $0 < \alpha_i \leq C$ (draw)
  - If $\gamma_i > 0$ (i.e., point within margin or misclassified), then $\alpha_i = C$ (draw)

Optimization

- $\ell_{w,b}(x,y) = \max\big(0, 1 - y(\langle w, x\rangle + b)\big)$
- Optimize loss function

$$L = \frac{C}{n}\sum_i \ell_{w,b}(x_i, y_i) + \frac{1}{2}\|w\|_2^2$$

  - Normalize by n this time, same problem just rescaled
  - Note similarity to ridge regression (C on former vs λ on latter)
- Gradient descent: $\frac{\partial L}{\partial w} = w + \frac{C}{n}\sum \delta_i$
  - $\delta_i = -y_i x_i$ if $1 - y_i\hat{y_i} \geq 0$, $\delta_i = 0$ if $1 - y_i\hat{y_i} \leq 0$ (draw, note non-diff pt)
- Could also run *projected* gradient descent on dual (draw, discuss)
- Solving dual using other methods are most practical