SVM and Complementary Slackness

Julia Kempe & David Rosenberg

CDS, NYU

February 19, 2019

SVM Review: Primal and Dual Formulations

The SVM Dual Problem

- We found the SVM dual problem can be written as:

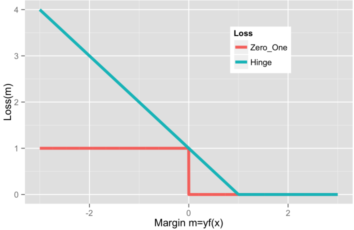
$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \ldots, n.
\end{aligned}
$$

- Given solution α^* to dual, primal solution is $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.
- Note $\alpha_i^* \in [0, \frac{c}{n}]$. So c controls max weight on each example. (**Robustness**!)

Insights From Complementary Slackness: Margin and Support Vectors

The Margin and Some Terminology

- For notational convenience, define $f^*(x) = x^T w^* + b^*$.
- Margin $yf^*(x)$



- Incorrect classification: $yf^*(x) \leqslant 0$.
- Margin error: $yf^*(x) < 1$.
- “On the margin”: $yf^*(x) = 1$.
- “Good side of the margin”: $yf^*(x) > 1$.

Support Vectors and The Margin

- Recall "**slack variable**" $\xi_i^* = \max(0, 1 - y_i f^*(x_i))$ is the hinge loss on (x_i, y_i).
- Suppose $\xi_i^* = 0$.
- Then $y_i f^*(x_i) \geqslant 1$
  - "on the margin" $(=1)$, or
  - "on the good side" (>1)

Complementary Slackness Conditions

- Recall our primal constraints and Lagrange multipliers:

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$(1 - y_i f(x_i)) - \xi_i \leqslant 0$

- Recall first order condition $\nabla_{\xi_i} L = 0$ gave us $\lambda_i^* = \frac{c}{n} - \alpha_i^*$.
- By strong duality, we must have **complementary slackness**:

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$
$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

Consequences of Complementary Slackness

- By strong duality, we must have **complementary slackness**:

$$\alpha_i^*\left(1-y_i f^*(x_i)-\xi_i^*\right)=0$$
$$\left(\frac{c}{n}-\alpha_i^*\right)\xi_i^*=0$$

- If $y_i f^*(x_i)>1$ then the margin loss is $\xi_i^*=0$, and we get $\alpha_i^*=0$.
- If $y_i f^*(x_i)<1$ then the margin loss is $\xi_i^*>0$, so $\alpha_i^*=\frac{c}{n}$.
- If $\alpha_i^*=0$, then $\xi_i^*=0$, which implies no loss, so $y_i f^*(x)\geqslant 1$.
- If $\alpha_i^*\in\left(0,\frac{c}{n}\right)$, then $\xi_i^*=0$, which implies $1-y_i f^*(x_i)=0$.

Complementary Slackness Results: Summary

$$
\begin{aligned}
\alpha_i^* = 0 \quad &\implies \quad y_i f^*(x_i) \geqslant 1 \\
\alpha_i^* \in \left(0, \frac{c}{n}\right) \quad &\implies \quad y_i f^*(x_i) = 1 \\
\alpha_i^* = \frac{c}{n} \quad &\implies \quad y_i f^*(x_i) \leqslant 1
\end{aligned}
$$

$$
\begin{aligned}
y_i f^*(x_i) < 1 \quad &\implies \quad \alpha_i^* = \frac{c}{n} \\
y_i f^*(x_i) = 1 \quad &\implies \quad \alpha_i^* \in \left[0, \frac{c}{n}\right] \\
y_i f^*(x_i) > 1 \quad &\implies \quad \alpha_i^* = 0
\end{aligned}
$$

Support Vectors

- If α^* is a solution to the dual problem, then primal solution is

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

  with $\alpha_i^* \in [0, \frac{c}{n}]$.
- The x_i's corresponding to $\alpha_i^* > 0$ are called **support vectors.**
- Few margin errors or "on the margin" examples $\implies$ **sparsity in input examples**.

Complementary Slackness To Get b^*

The Bias Term: b

- For our SVM primal, the complementary slackness conditions are:

$$\alpha_i^* \left(1 - y_i \left[x_i^T w^* + b\right] - \xi_i^*\right) = 0 \quad (1)$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0 \quad (2)$$

- Suppose there's an i such that $\alpha_i^* \in \left(0, \frac{c}{n}\right)$.
- (2) implies $\xi_i^* = 0$.
- (1) implies

$$\begin{aligned} & y_i \left[x_i^T w^* + b^*\right] = 1 \\ \iff \quad & x_i^T w^* + b^* = y_i \text{ (use } y_i \in \{-1, 1\}) \\ \iff \quad & \boxed{b^* = y_i - x_i^T w^*} \end{aligned}$$

The Bias Term: b

- The optimal b is
$$b^* = y_i - x_i^T w^*$$
- We get the same b^* for any choice of i with $\alpha_i^* \in \left(0, \frac{c}{n}\right)$
  - **With exact calculations!**
- With numerical error, more robust to average over all eligible i's:
$$b^* = \text{mean}\left\{ y_i - x_i^T w^* \mid \alpha_i^* \in \left(0, \frac{c}{n}\right) \right\}.$$
- If there are no $\alpha_i^* \in \left(0, \frac{c}{n}\right)$?
  - Then we have a **degenerate SVM training problem**[1] ($w^* = 0$).

[1] See Rifkin et al.'s "A Note on Support Vector Machine Degeneracy", an MIT AI Lab Technical Report.

Teaser for Kernelization

Dual Problem: Dependence on x through inner products

- SVM Dual Problem:

$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \quad i = 1, \ldots, n.
\end{aligned}
$$

- Note that all dependence on inputs x_i and x_j is through their inner product: $\langle x_j, x_i \rangle = x_j^T x_i$.
- We can replace $x_j^T x_i$ by any other product...
- This is a "kernelized" objective function.