Support Vector Machines

What Is a Support Vector Machine?

1. A subset of the training examples \( x \) (the support vectors)
2. A vector of weights for them \( \alpha \)
3. A similarity function \( K(x,x') \) (the kernel)

Class prediction for new example \( x_q \):

\[
f(x_q) = \text{sign}
\left( \sum_i \alpha_i y_i K(x_q, x_i) \right)
\]

\( \{y_i \in \{-1,1\} \)
Examples of Kernels

- **Linear**: $K(x, x') = x \cdot x'$
- **Polynomial**: $K(x, x') = (x \cdot x')^d$
- **Gaussian**: $K(x, x') = \exp(-\frac{1}{2\sigma^2}\|x - x'\|^2)$

Learning SVMs

- How do we:
  - Choose the kernel? Black art
  - Choose the examples? Side effect of choosing weights
  - Choose the weights? Maximize the margin

The Weight Optimization Problem

- **Margin** = $\min y_i(w \cdot x_i)$
- Easy to increase margin by increasing weights
- Instead: Fix margin, minimize weights
- **Minimize** $\frac{1}{2}w \cdot w$, $y_i(w \cdot x_i) \geq 1$, for all $i$

Example: Polynomial Kernel

$u = (u_1, u_2)$
$v = (v_1, v_2)$

$\langle u \cdot v \rangle = (u_1v_1 + u_2v_2)^2$

$= u_1^2v_1^2 + u_2^2v_2^2 + 2u_1v_1u_2v_2$

$= (u_1^2, u_2^2, \sqrt{2}u_1u_2, \sqrt{2}u_1u_2)$

$= \phi(u) \cdot \phi(v)$

- Linear kernel can’t represent quadratic frontiers
- Polynomial kernel can

Maximizing the Margin

Constrained Optimization 101

- Minimize $f(w)$
- Subject to $h_i(w) = 0$, for $i = 1, 2, \ldots$
- At solution $w^*$, $\nabla f(w^*)$ must lie in subspace spanned by $\{\nabla h_i(w^*): i = 1, 2, \ldots\}$
- **Lagrangian function**:

$L(w, \beta) = f(w) + \sum \beta_i h_i(w)$

- The $\beta_i$ are the Lagrange multipliers
- Solve $\nabla L(w^*, \beta) = 0$
Primal and Dual Problems

- Problem over $w$ is the **primal**
- Solve equations for $w$ and substitute
- Resulting problem over $\beta$ is the **dual**
- If it’s easier, solve dual instead of primal
- In SVMs:
  
  - Primal problem is over feature weights
  - Dual problem is over instance weights

Solution Techniques

- Use generic quadratic programming solver
- Use specialized optimization algorithm
- E.g., SMO (Sequential Minimal Optimization)
  
  - Simplest method: Update one $\alpha_i$ at a time
  - But this violates constraints
  
  - Iterate until convergence:
    1. Find example $x_i$ that violates KKT conditions
    2. Select second example $x_j$ heuristically
    3. Jointly optimize $\alpha_i$ and $\alpha_j$

Handling Noisy Data

- Introduce slack variables $\zeta_i$
- Minimize $\frac{1}{N} \sum_{i=1}^{N} (y_i(w \cdot x_i) - 1)^+ + \sum_{i=1}^{N} \xi_i$ for all $i$
Support Vector Machines: Summary

- What is a support vector machine?
- The perceptron revisited
- Kernels
- Weight optimization
- Handling noisy data