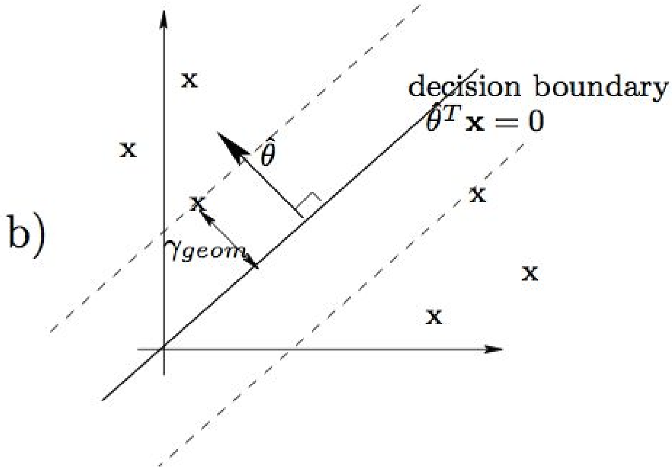
Machine Learning

More on the Support Vector Machine

Maximum Margin



- **Find the maximum margin linear classifier**
  - Identify any classifier any classifier that correctly classifies all the examples.
  - Increase the margin till reach the extreme
  - The solution is unique

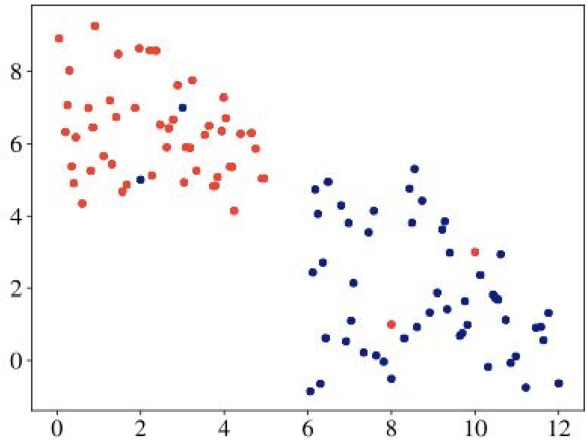
- **To find the optimal theta:**
  - $$\text{minimize } \frac{1}{2}\|\theta\|^2/\gamma^2 \text{ subject to } y_t\theta^T\mathbf{x}_t \geq \gamma \text{ for all } t = 1, \ldots, n$$
    - $\min \frac{1}{2}||\mathbf{w}||^2$, such that $y_i(\mathbf{x}_i\mathbf{w} - b) - 1 \geq 0, i = 1, \ldots, N.$ in the Hundred Page Machine Leaning Book).
  - We only care about the ratio of theta and gamma. So let gamma be 1. Get get a simplified problem (the **standard SVM form**) :

$$\text{minimize } \frac{1}{2}\|\theta\|^2 \text{ subject to } y_t\theta^T\mathbf{x}_t \geq 1 \text{ for all } t = 1, \ldots, n$$

- The resulting geometric margin is 1/||^theta||, where ^theta is the unique solution to the problem.

Dealing with Noise

- Noise: Makes the data not linearly separable => Some outliers that make SVM unable to find a line which perfectly separate the positive example from the negative ones.

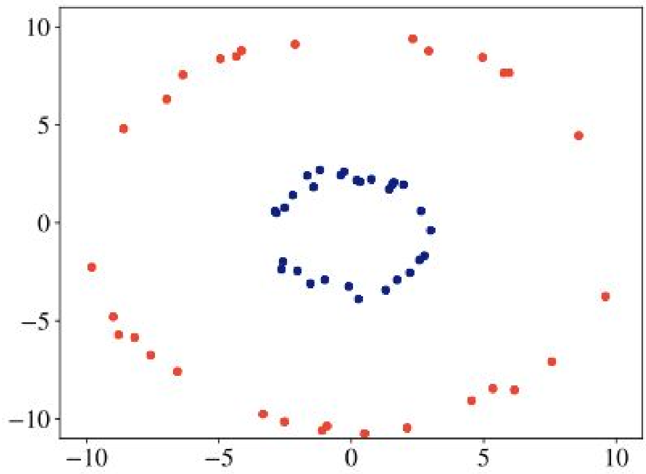


- The **Hinge Loss** Function
  - $\max(0, 1 - y_i(\mathbf{w}\mathbf{x}_i - b))$
  - If the predicted value lies on the right side, the function = 0
  - If on the wrong side, the function's value is proportional to the distance from the decision boundary.
- The optimal solution will be the case when the misclassification and maximizing margin are balanced.
  - A cost function for misclassification and increasing margin:

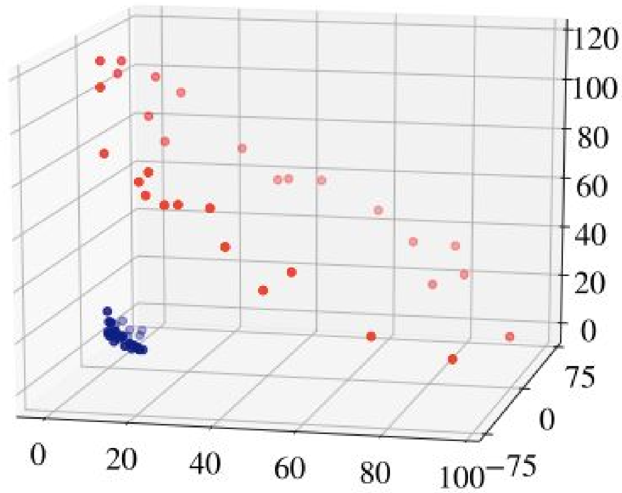
$$C\|\mathbf{w}\|^2 + \frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - y_i(\mathbf{w}\mathbf{x}_i - b)),$$

    - C ::= Hyperparameter
      - "Determines the tradeoff btw increasing the margin and ensuring each x lies on the right side."
    - Usually chosen experimentally
    - As C gets greater, the cost for misclassification (the second term of the function) becomes negligible. => SVM will try to find the highest margin by ignoring misclassification.

Dealing with Inherent Non-Linearity



SVM can work with non-linear data by transforming the data into a higher dimension:



- To increase dimension: map, with specific mapping function which we don't know in priori.
    - In the previous example: mapping function $\phi([q,p]) \stackrel{\text{def}}{=} (q^2, \sqrt{2}qp, p^2)$
- The **Kernel Trick:**
    - "Using a function to implicitly transform the original space into a higher dimensional space during the cost function optimization".
    - Solves the problem that we don't know which mapping function would work in priori.
    - More on the summary of Kernel.