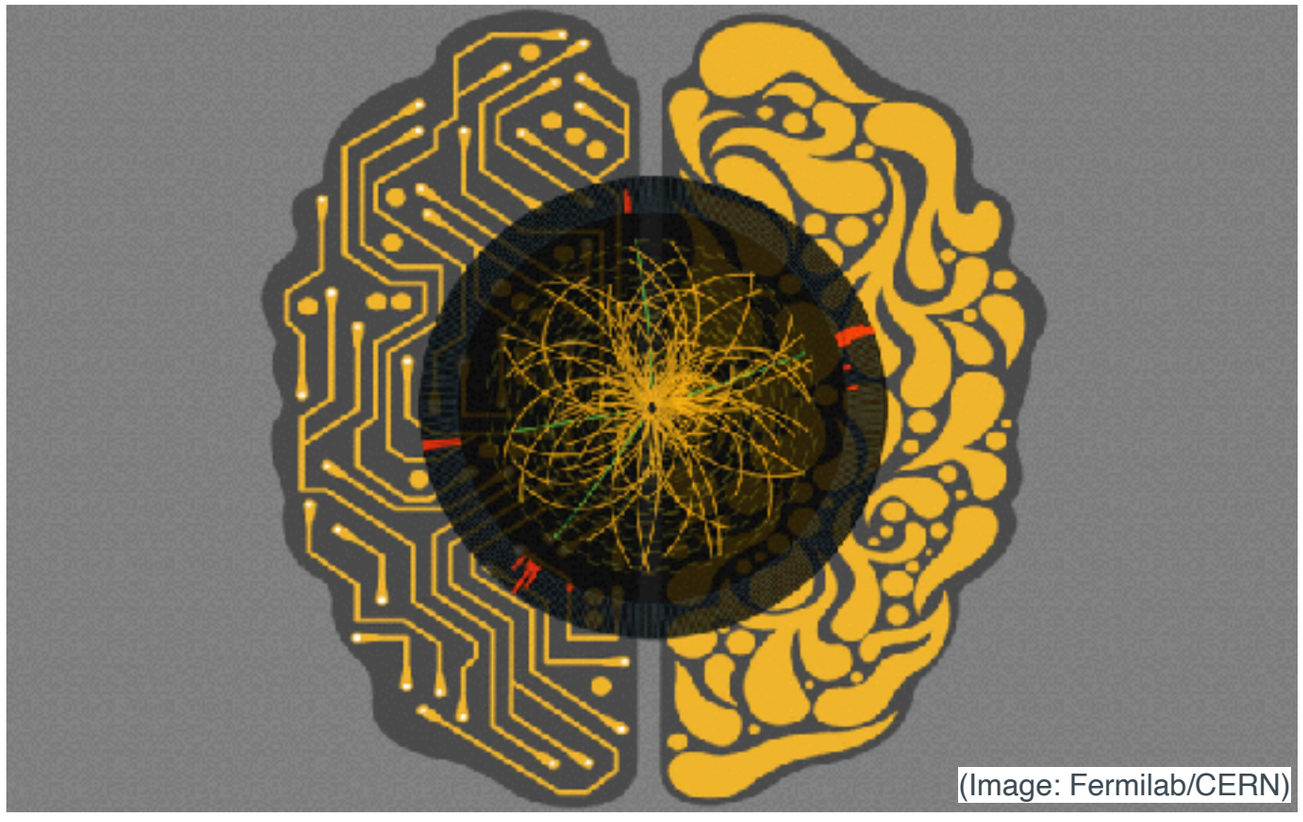
PHY 835: Collider Physics Phenomenology

Machine Learning in Fundamental Physics

Gary Shiu, UW-Madison



Lecture 8: Support Vector Machine (SVM)

Recap of Lecture 7

- Capacity of Perceptron = 2 bits/weight
- Survey of classifiers
- Decision Trees
- How can we use decision tree for regression?

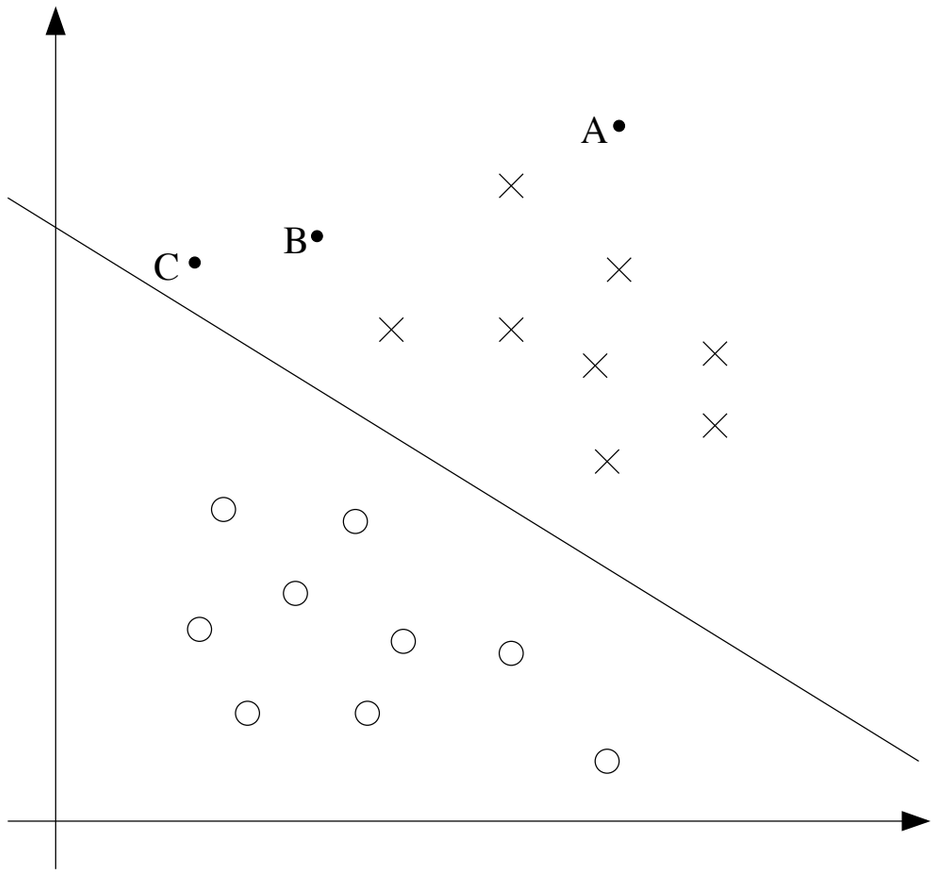
Outline for today

- Support Vector Machines: functional and geometric margins
- Optimal margin classifier
- Lagrange duality
- Kernel Methods

Ref: Andrew Ng's Lecture Notes: https://sgfin.github.io/files/notes/CS229_Lecture_Notes.pdf

Support Vector Machines

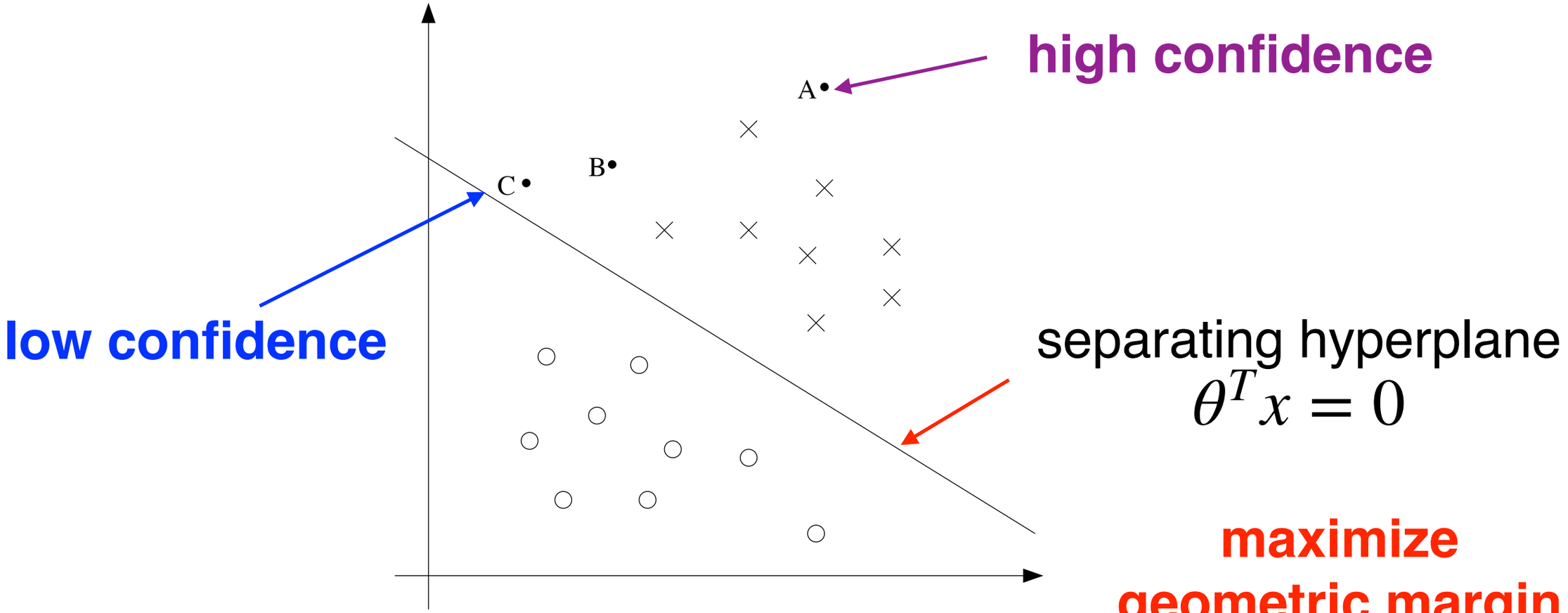
- Among the best off-the-shelf **supervised learning** algorithms.
- The idea of **margins**: separating data with a large **"gap"**.
- Start with linear separable patterns.



- Kernel methods allow us to generalize this to non-linear separation.

Margins: Intuition

- Logistic regression: $h_\theta = g(\theta^T x) = 1$ if $\theta^T x \geq 0$ and 0 otherwise.
- High confidence if $\exists \theta$ s.t. $\theta^T x \gg 0$ or $\theta^T x \ll 0$ in the training set: **large functional margin**.
- Decision boundary (or **separating hyperplane**):



Functional Margins

- Use $y \in \{-1,1\}$ instead of $\{0,1\}$ as binary output.
- The linear classifier:

$$h_{w,b}(x) = g(w^T x + b)\ ; \qquad g(z) = 1 \text{ if } z \geq 0 \text{ and } -1 \text{ otherwise}$$

- Define the **functional margin** of a training sample:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$

- To maximize the functional margin, $w^T x + b \gg 0$ if $y^{(i)} = 1$ and $w^T x + b \ll 0$ if $y^{(i)} = -1$.
- Large functional margin represents a **confident** and a **correct** prediction.

Functional Margins

- However, an undesirable feature of this measure is that it is not invariant under rescaling, e.g.,:

$$w \to 2w \ , \quad b \to 2b \ , \quad g(w^T x + b) = g(2w^T x + 2b)$$

- This increases the functional margin but this rescaling should not change the decision $h_{w,b}(x)$.

- Seems sensible to impose some normalization condition, e.g., $||w||_2 = 1$ and

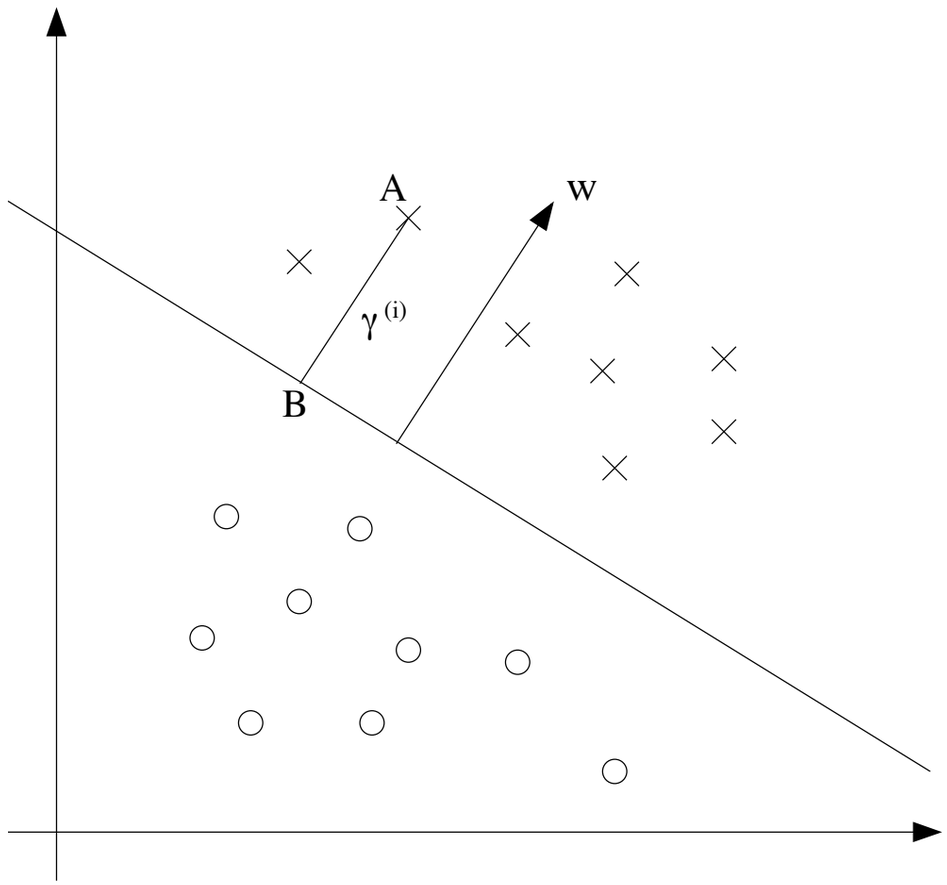
$$(w, b) \to (w/||w||_2, b/||w||_2)$$

- Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1, \ldots, n\}$, define the functional margin wrt the set S:

$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}$$

Geometric Margins

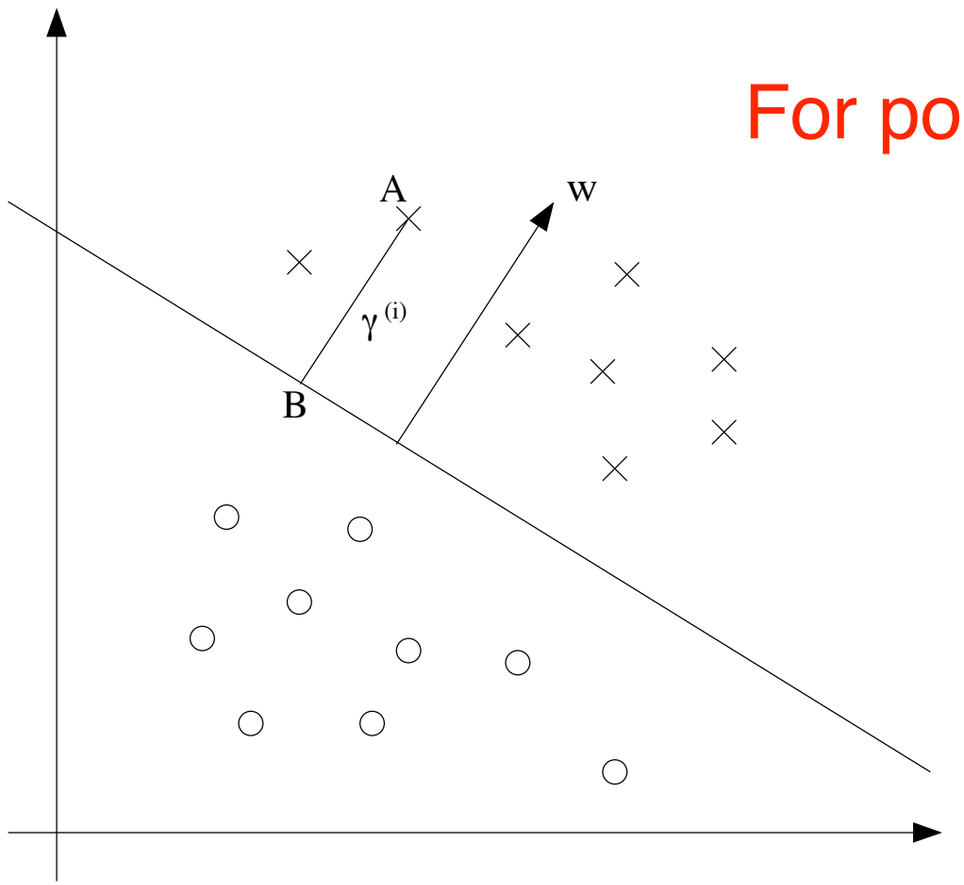
- w is orthogonal to the separating hyperplane (convince yourself):



- Maximizing the margin means maximizing the distance $\gamma^{(i)}$ to the decision boundary.

Geometric Margins

- Point B is given by $x^{(i)} - \gamma^{(i)} \cdot w/||w||$



For point B to lie on the decision boundary:

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

$$\Rightarrow \quad \gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||}.$$

- The above was derived for a positive training sample. Generally:

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

Geometric Margins

- Comparing the functional margin with the geometric margin:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b) \qquad \gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}\right)$$

- They coincide if $||w|| = 1$ but the geometric margin has the advantage that it is invariant under arbitrary rescaling:

$$||W|| = 1 \text{ , or } |w|_1 = 5 \text{ , or } |w_1 + b| + |w_2| = 2$$

- The geometric margin wrt to a training set S is:

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)}$$

Optimal Margin Classifier

- Find a decision boundary that **maximizes the margin** → a classifier that separates the positive/negative training samples with a **gap**.
- Optimization problem:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

- Each training sample has functional margin at least γ ; moreover, the $||w|| = 1$ constraint ensures functional margin=geometric margin.
- The $||w|| = 1$ makes this a nasty minimization problem. Consider instead:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \frac{\hat{\gamma}}{||w||} \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m \end{aligned}$$

still difficult, non-convex optimization problem

Optimal Margin Classifier

- Instead, we can apply a rescaling of w and b to bring

$$\hat{\gamma} = 1$$

- The optimization problem can be turned into:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

one with a convex quadratic objective and only linear constraints; can be solved using commercial quadratic programming (QP) code.

- The solution is the **optimal margin classifier**.
- **Lagrange duality** (next) allows us to use kernels to get optimal margin classifiers to work efficiently in a very high dimensional spaces.

Lagrange Duality

- Consider solving constrained optimization problems of the form:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

- Define the **Lagrangian**:

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

- The optimized solution is given by:

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0$$

- This can be generalized to constrained optimization problems with inequalities as well. Consider this **primal optimization problem**:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Lagrange Duality

- Define the **generalized Lagrangian**:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

- Consider the quantity:

$$\theta_{\mathcal{P}}(w) = \max_{\alpha, \beta \,:\, \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta).$$

- If w violates any of the primal constraints:

$$\begin{aligned} \theta_{\mathcal{P}}(w) &= \max_{\alpha, \beta \,:\, \alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) \\ &= \infty. \end{aligned}$$

- Conversely, if the constraints are satisfied, $\theta_P(w) = f(w)$, therefore:

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

Lagrange Duality

- The minimization problem is equivalent to the original problem:

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta),$$

- The optimal value of the objective = the value of the primal problem:

$$p^* = \min_w \theta_{\mathcal{P}}(w)$$

- Define the dual:

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

- The **dual optimization problem**:

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta).$$

Lagrange Duality

- The optimal value of the dual problem's objective is:

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(w)$$

- Using the fact that "max min" of a function ≤ its "min max":

$$d^* = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

- Under some conditions (known as the KKT conditions, given shortly):

$$d^* = p^*$$

- We can solve the dual problem in lieu of the primal problem.

KKT Conditions

- Suppose f and the g_i's are **convex**, and the h_i's are **affine** and further that the constraints g_i's are strictly feasible, meaning $\exists w$ s.t.

$$g_i(w) < 0 \quad \forall i$$

- Under these assumptions:
  - there must $\exists w^*, \alpha^*, \beta^*$ such that w^* solves the primal problem, α^*, β^* solve the dual problem, and

$$p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$$

- w^*, α^*, β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

KKT Conditions

- If some w^*, α^*, β^* satisfy the KKT conditions:

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

then it is also a solution to the primal and dual problems.

- The middle equation is known as the **dual complementarity condition** which implies that if $\alpha_i^* > 0$, then $g_i(w) = 0$.

- This property is key to showing that SVM has only a small number of **support vectors**.

Finding Optimal Margin Classifiers

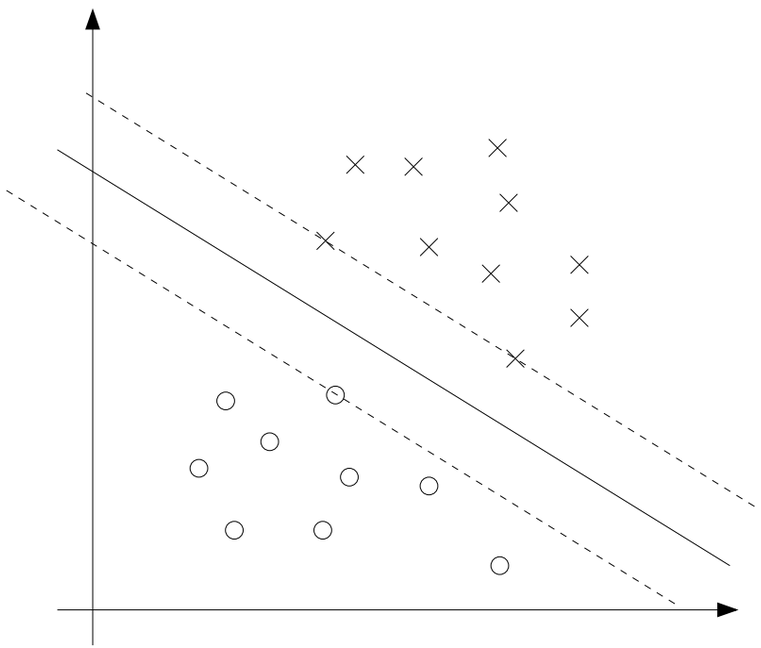
- Recall our primal optimization problem:

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

- We can write the constraints as (one for each training sample):

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0.$$

- The KKT dual complementarity condition implies that $\alpha_i > 0$ only for the training samples that have functional margin =1



only three $\alpha_i > 0$, the corresponding three points are called the support vectors; # support vectors can be << $||S||$

Finding Optimal Margin Classifiers

- Develop the dual problem, and express the algorithm in terms of $< x^{(i)}, x^{(j)} >$ between points in the input feature space (kernel trick).
- The Lagrangian for our optimization problem has only α_i but no β_i.

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

- To find θ_D:

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \qquad \frac{\partial}{\partial b} \mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- Plugging back into the Lagrangian:

$$\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - b \cancel{\sum_{i=1}^{m} \alpha_i y^{(i)}}$$

Finding Optimal Margin Classifiers

- The dual optimization problem:

$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \end{aligned}$$

- Check that the KKT conditions are satisfied, so we can solve the dual problem in lieu of the primal problem:

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \qquad b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}$$

- After fitting our model parameters to a training set, make a prediction at a new input point x depending on the value of:

$$w^T x + b = \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b = \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.$$

depends on inner products, and only with support vectors

Summary

- Support Vector Machines: functional and geometric margins
- Optimal margin classifier
- Lagrange duality
- Kernel Methods