Lecture 7 | Machine Learning (Stanford)

- Assume the problem is **linearly separable**.
- **Optimal margin classifier**: maximizing geometric margin
  - final form: #3 $$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$ $$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$
  - convex quadratic programming problem, could be solved inefficiently.
- **Lagrange Duality**
  - d* = p* when satisfying KKT conditions
- #3 -> Lagrangian (8) -> dual form of the problem -> solve alpha -> get w -> get b

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

Lecture 8 | Machine Learning (Stanford)

- For **linearly non-separable** problem: mapping into higher dimensions
- **Kernels**: help to compute **<x(i), x(j)>** efficiently, even if x is very high dimensional, $O(n^2)$ --> $O(n)$
  - Kernels can be applied to any machine learning algorithms, as long as there's **<x(i), x(j)>** in expressions.
- **Mercer Theorem**: a kernel function is valid iff kernel matrix is symmetric positive semi-definite.
- **Soft margin**: make the classifier less sensitive to outliers
- **SMO**: to solve the dual problem (take use of **coordinate ascent**)