MATH 565: Lecture 21 (03/31/2026)

Today: * KKT for Hinge-loss SVM

Recall KKT conditions for $\min\{F(\bar{w}) \mid f_i(\bar{w}) \le 0, i=1,\dots,m\}$ (P)

$$L_{\bar{\alpha}} = \min_{\bar{w}} F(\bar{w}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{w}), \quad \bar{\alpha} \ge \bar{0}$$

$$\max_{\bar{\alpha} \ge \bar{0}} L_{\bar{\alpha}} = \max_{\bar{\alpha} \ge \bar{0}} \min_{\bar{w}} \{ \underbrace{F(\bar{w}) + \sum_{i=1}^{m} \alpha_i f_i(\bar{w})}_{H(\bar{w}, \bar{\alpha})} \} \quad \text{(D)}$$

For convex $F(\cdot)$ and convex $f_i(\cdot)$, $(\bar{w}, \bar{\alpha})$ is optimal for (P) and (D) iff

* feasibility: $f_i(\bar{w}) \le 0 \ \forall i$, $\alpha_i \ge 0 \ \forall i$
* CSCs: $\alpha_i f_i(\bar{w}) = 0 \ \forall i$
* Stationarity: $\nabla_{\bar{w}} F(\bar{w}) + \sum_{i=1}^{m} \alpha_i \nabla_{\bar{w}} f_i(\bar{w}) = \bar{0}$

} KKT conditions

The Stationarity constraints are also called as primal-dual (PD) constraints, and are usually used to eliminate the primal variables to create a maximization problem only in the dual variables with a much simpler set of constraints, e.g., box constraints.

Recall that in the case of equality constraints, these constraints were used to eliminate (a subset of) variables.

We now describe in detail how the KKT conditions can be used to efficiently solve ML problems, with the case of hinge-loss SVM as a specific application case.

Hinge Loss SVM using KKT Conditions

Recall the loss function minimization for L_1-SVM (hinge loss SVM):

$$\min \; J = \sum_{i=1}^{n} \max\{0, 1 - y_i(\bar{w}^T \bar{x}_i)\} + \frac{\lambda}{2}\|\bar{w}\|^2$$

Equivalently, we write

$$\min \; J = C \sum_{i=1}^{n} \max\{0, 1 - y_i(\bar{w}^T \bar{x}_i)\} + \frac{1}{2}\|\bar{w}\|^2$$

→ this setting is more standard for SVM problems

for $C = \frac{1}{\lambda} > 0$, the slack penalty parameter.

Using ξ_i for the slack term, the primal problem becomes

$$\min \; J = \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

$$(P) \quad \text{s.t.} \quad \xi_i \geq 1 - y_i(\bar{w}^T \bar{x}_i), \quad i = 1, \ldots, n \;\rightarrow\; \text{margin constraints}$$

$$\xi_i \geq 0, \quad i = 1, \ldots, n \;\rightarrow\; \text{non negativity}$$

The Lagrangian relaxation of (P) is

$$\min \; L(\bar{\alpha}, \bar{r}) = \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \left(\xi_i - 1 + y_i(\bar{w}^T \bar{x}_i)\right) - \sum_{i=1}^{n} r_i \xi_i$$

relaxing margin constraints; relaxing ≥ 0

The dual problem becomes

$$L^* = \max_{\bar{\alpha}, \bar{r} \geq \bar{0}} \; \min_{\bar{w}, \xi} \; L(\bar{\alpha}, \bar{r}) \quad (D)$$

This is a max-min problem. Let's apply KKT conditions to solve (D).

Since the constraints in (P) are $\geq$, we have $-\alpha_i$ and $-r_i$ as the multipliers in the Lagrangian L (for each of the two sets of constraints). Recall that the penalty terms are written as $\sum \alpha_i f_i(\bar{w})$ for $f_i(\bar{w}) \leq \bar{0}$ constraints.

ξ_i terms

$$L(\bar{\alpha},\bar{r}) = \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(\xi_i - 1 + y_i(\bar{w}^T\bar{x}_i)\right) - \sum_{i=1}^{n} r_i\xi_i$$

Stationarity Conditions: $\nabla_{\bar{w}} L = \bar{0}$ and $\nabla_{\bar{\xi}} L = \bar{0} \equiv \frac{\partial L}{\partial \xi_i} = 0$

$$\Rightarrow \quad \bar{w} - \sum_{i=1}^{n}\alpha_i y_i \bar{x}_i = \bar{0} \qquad \text{——— (1)}$$

and

$$C - \alpha_i - r_i = 0 \;\forall i \qquad \text{——— (2)}$$

We can use (1) to eliminate $\bar{w}$. But how about ξ_i?

Note that the coefficient of ξ_i in L is $(C - \alpha_i - r_i)$.

Hence, by (2), we can drop ξ_i from L at optimality!

$$\Rightarrow L(\bar{w},\bar{\alpha}) = \frac{1}{2}\|\bar{w}\|^2 + \sum_{i=1}^{n}\alpha_i\left(1 - y_i(\bar{w}^T\bar{x}_i)\right)$$

→ note: $\bar{r}$ has disappeared!

So, (1) $\Rightarrow$

$$L(\bar{\alpha}) = \frac{1}{2}\left\|\sum_{i=1}^{n}\alpha_i y_i \bar{x}_i\right\|^2 + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\sum_{j=1}^{n}\alpha_j y_j \bar{x}_j^T\bar{x}_i\right)$$

$$= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \bar{x}_i^T\bar{x}_j$$

r_i is not present here, but we still need $r_i \geq 0 \;\forall i$.

So, (2) $\Rightarrow$ $r_i = C - \alpha_i \geq 0$

$\Rightarrow$ We need $0 \leq \alpha_i \leq C$ for dual variables, i.e., they are box constraints!

We can now write the dual problem (D) as an equivalent minimization problem with box constraints.

$$\min_{0 \le \bar{\alpha} \le \bar{C}} L_D = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \bar{x}_i^T \bar{x}_j - \sum_{i=1}^{n} \alpha_i \qquad (D)$$

It can be shown that L_D is always convex. The leading term is $\bar{\alpha}^T H \bar{\alpha}$ for $H \succeq 0$, the PSD matrix of similarities between $\bar{x}_i$'s.

$\Rightarrow$ (D) is a convex optimization problem!

In fact,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \bar{x}_i^T \bar{x}_j = \bar{\alpha}^T B B^T \bar{\alpha}$$

where $B = \bar{y} X$, i.e., $B_{i\cdot} = y_i \bar{x}_i^T$, the i^{th} row of B

data instance (or vector) scaled by its y_i ($= \pm 1$) value.

We can use the methods we saw previously to solve (D), e.g., gradient descent with box projection. We can then extract the solution to (P) directly from the solution to (D).

(1) gives $\bar{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \bar{x}_i$, and we can plug in $\bar{w}^*$ values into (P) constraints to get ξ_i^* values.

$$\xi_i^* = \max\{0, 1 - y_i(\bar{w}^{*T} \bar{x}_i)\}.$$

Solving (D)

Using Gradient Descent

Recall the dual problem:

$$\min L_D = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \bar{x}_i^T \bar{x}_j - \sum_{i=1}^{n} \alpha_i \qquad (D)$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C \quad \forall i.$$

The first order optimality conditions become

$$\frac{\partial L_D}{\partial \alpha_k} = y_k \sum_{l=1}^{n} y_l \alpha_l \bar{x}_k^T \bar{x}_l - 1 = 0.$$

We can do the GD update as follows:

$$\bar{\alpha} \leftarrow \bar{\alpha} - \eta \left[\frac{\partial L_D}{\partial \bar{\alpha}} \right]$$

($\nabla_{\bar{\alpha}} L_D$; η → learning rate)

And if any component of $\bar{\alpha}$ is outside the box $[0, C]$, then we simply project it to the box boundary. In details,

if $\alpha_i < 0$, $\alpha_i \to 0$;

else if $\alpha_i > C$, $\alpha_i \to C$.