MATH 565: Lecture 3 (01/20/2026)

Today:
* more on SVM
* Taylor expansion
* local optimality conditions

Recall: SVM model:

$$\min_{\bar{w},b,\bar{\xi}} \; J = C\sum_{i=1}^{n}\xi_i + \frac{1}{2}\|\bar{w}\|^2 \qquad (C>0)$$

$$\text{s.t.} \quad y_i(\bar{w}^T\bar{x}_i + b) \geq 1-\xi_i, \quad i=1,\dots,n$$

$$\xi_i \geq 0 \quad \forall i$$

Let's examine the unified (main) constraints in detail:

Recall: the model tries to get $\bar{w}^T\bar{x}_i+b\geq 1$ when $y_i=+1$ and $\bar{w}^T\bar{x}_i+b\leq -1$ when $y_i=-1$.

ξ_i: measures by how much the i^{th} sample violates well-separatedness

When $y_i=+1$, we get $\bar{w}^T\bar{x}_i+b\geq 1-\xi_i$

For instance, if $\bar{w}^T\bar{x}_i+b=0.7$, then $\xi_i\geq 0.3$ is needed;
but if $\bar{w}^T\bar{x}_i+b=2$, then $\xi_i=0$ works, and will be set to this value because of the $C\xi_i$ term in the objective function (recall, $\xi_i\geq 0$).

On the other hand, if $y_i=-1$, the constraint becomes

$$\bar{w}^T\bar{x}_i+b\leq -1+\xi_i.$$

For instance, if $\bar{w}^T\bar{x}_i+b=-3$, $\xi_i=0$ in the optimal solution.
But if $\bar{w}^T\bar{x}_i+b=0.5$, we need $\xi_i\geq 1.5$ for the constraint to hold.

Technically, we should be including the affine variable b in the regularizing term:

affine variable = or intercept

$$\min_{\bar{w}, b, \bar{\xi}} \; J = C \sum_{i=1}^{n} \xi_i + \frac{1}{2} \left\| \begin{bmatrix} \bar{w} \\ b \end{bmatrix} \right\|^2$$

$$\text{s.t.} \quad y_i (\bar{w}^T \bar{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0 \;\; \forall i$$

Another equivalent notation is to write $\bar{w} = [w_0 \; w_1 \ldots w_d]^T$, and write the model as

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

$$\text{s.t.} \quad y_i \left(\bar{w}^T \begin{bmatrix} 1 \\ \bar{x}_i \end{bmatrix} \right) \geq 1 - \xi_i, \quad i = 1, \ldots, n,$$

$$\xi_i \geq 0 \;\; \forall i.$$

Note how the well-separated instances do not contribute to the objective function — irrespective of the extent of their well-separatedness. At the same time, instances that are not well-separated do incur a loss, i.e., they add to the loss function. Furthermore, the amount of this loss is more when the instance violates well-separatedness more.

We will study SVM models in detail later on...

Taylor Expansion

In one dimension, the Taylor expansion of $f(x)$ at $x=a$ is given by

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a) + \cdots + \frac{(x-a)^r}{r!}\frac{d^r f(x)}{dx^r}\bigg|_{x=a} + \cdots$$

When $|x-a|$ is small, we can take

$$f(x) \approx f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2!}f''(a)$$

as a reasonable and accurate representation of $f(x)$.

In d-dimensions, the Taylor expansion of $f(\bar{x})$ at $\bar{x}=\bar{a}$ is

$$f(\bar{x}) = f(\bar{a}) + \sum_{i=1}^{d}(x_i - a_i)\left[\frac{\partial f}{\partial x_i}\right]\bigg|_{\bar{x}=\bar{a}} + \sum_i\sum_j \frac{(x_i-a_i)(x_j-a_j)}{2!}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]\bigg|_{\bar{x}=\bar{a}} + \cdots$$

Equivalently,

$$f(\bar{x}) = f(\bar{a}) + [\bar{x}-\bar{a}]^T \nabla f(\bar{a}) + [\bar{x}-\bar{a}]^T Hf(\bar{a})[\bar{x}-\bar{a}] + \cdots$$

for gradient $\nabla f(\bar{a})$ and Hessian $Hf(\bar{a})$ of $f(\bar{x})$ at $\bar{x}=\bar{a}$.

We will employ Taylor series approximations of functions for multiple purposes in this class – both for devising efficient algorithms and for deriving theoretical results, i.e., proofs. Despite its somewhat simple form, the Taylor series expansion proves to be a powerful tool!

We now present local optimality conditions first in 1D and then in d-dimensions in general. We will use the Taylor series expansion to justify them.

Local Optimality Conditions in 1D

Lemma 1 Let $f: \mathbb{R} \to \mathbb{R}$ be a function. Then $f(x)$ is a minimum value at $x = x_0$ with respect to its immediate locality if

$f'(x_0) = 0$ and → (first order optimality condition)

$f''(x_0) > 0$ → (second order optimality condition)

in a small enough neighborhood of $x = x_0$

Proof Consider the Taylor expansion of f at x_0 for $x = x_0 + \Delta$: $(\Delta \in \mathbb{R})$

$$f(x_0+\Delta) \approx f(x_0) + \Delta f'(x_0) + \frac{\Delta^2}{2} f''(x_0)$$

($\Delta f'(x_0) = 0$, $\frac{\Delta^2}{2} f''(x_0) > 0$)

We can take $|\Delta| \ll 1$ small enough such that

$$\left|\frac{\Delta^2}{2} f''(x_0)\right| \gg \left|\sum \frac{\Delta^r}{r!} f^{(r)}(x_0)\right|$$

→ the second order term dominates all other higher order terms taken together

$\Rightarrow$ under first and second order optimality conditions,

$$f(x_0+\Delta) > f(x_0) \quad \text{for } |\Delta| \text{ small enough} \qquad \square$$

The first order condition ($f'(x) = 0$) is typically solved using **gradient descent**.

(initialization) Step 0. Start $x = x_0$ (can be chosen randomly)

do

Step k. $x_k \leftarrow x_{k-1} - \alpha f'(x_{k-1})$ (in general, $x \leftarrow x - \alpha f'(x)$)

while $|x_k - x_{k-1}| > \epsilon$.

Here $\alpha > 0$ is the learning rate (also called the step size) and $\epsilon \geq 0$ is a convergence tolerance.

Note that we are changing x by $\delta x = -\alpha f'(x)$. In fact, we are changing x along the "steepest descent" direction — which is trivial in 1D as we have only two options (↑ or ↓), but is nontrivial in d dimensions ($d \geq 2$) when we could have infinitely many options.

Our goal is to decrease $f(x)$ (we are minimizing it). Each step of gradient descent is guaranteed to decrease $f(x)$ for small values of α, since we have by Taylor expansion:

$$\begin{aligned} f(x+\delta x) &\approx f(x) + \delta x f'(x) \qquad \text{for } |\delta x| << 1. \\ &= f(x) - \alpha f'(x) \cdot f'(x) \\ &= f(x) - \alpha [f'(x)]^2 \\ &< f(x) \end{aligned}$$

By the same argument, we get that $f(x+\delta x) \approx f(x)$ when $f'(x) = 0$, which indicates we have converged to a local optimum.

Example

Consider $f(x) = x^2 \sin x + x$.

$$\Rightarrow \quad f'(x) = 2x\sin(x) + x^2\cos(x) + 1$$

and $$f''(x) = (2 - x^2)\sin(x) + 4x\cos(x)$$

We explore steps of gradient descent starting at $x_0 = 2$ and then at $x_0 = 5$ using $\alpha = 0.05$.

We get faster convergence for $x_0 = 5$ to a local minimum at $x^* = 5.05$.

See course web page for Python notebook...

Local Optimality Conditions in d Dimensions

Notation We use $\bar{w}, \bar{\xi}$ as variables ($\bar{w} = [w_1, \dots w_d]^T$ or $\bar{w} = [w_0, w_1, \dots w_d]^T$) while $\bar{x}, \bar{y}$ are data in our settings of optimization for ML. The objective function is a loss function typically denoted J, e.g., $J = \frac{1}{2}\|D\bar{w} - \bar{y}\|^2 + \frac{1}{2}\|\bar{w}\|^2$ for regularized regression.

Lemma 2 Let $J: \mathbb{R}^d \to \mathbb{R}$ be a loss function. Then $J(\bar{w})$ is a minimum value at $\bar{w} = \bar{w}_0$ with respect to its immediate locality if

$$\nabla J(\bar{w}_0) = \bar{0}, \text{ i.e., } \left[\frac{\partial J}{\partial w_1} \cdots \frac{\partial J}{\partial w_d}\right]^T \Bigg|_{\bar{w}=\bar{w}_0} = \bar{0}, \text{ and}$$

(first order optimality condition)

and $HJ(\bar{w}_0) \succ 0$, i.e., the Hessian at $\bar{w} = \bar{w}_0$ is positive definite,

(i.e., $\bar{w}^T H \bar{w} \geq 0 \;\; \forall \bar{w} \in \mathbb{R}^d \setminus \{\bar{0}\}$.

(second order optimality condition)

Similar to the 1D case, we can understand these conditions using the Taylor expansion of J at $\bar{w}_0$ with $\bar{w} = \bar{w}_0 + \epsilon\bar{v}$ for $\epsilon > 0$:

$$J(\bar{w}_0 + \epsilon\bar{v}) \approx J(\bar{w}_0) + \epsilon \bar{v}^T \underbrace{\nabla J(\bar{w}_0)}_{=\bar{0}} + \frac{\epsilon^2}{2} \bar{v}^T \underbrace{[HJ(\bar{w}_0)]\bar{v}}_{>0 \;\forall \bar{v} \neq \bar{0}}$$

Under the first and second order optimality conditions, we get $J(\bar{w}_0 + \epsilon\bar{v}) > J(\bar{w}_0) \;\; \forall \bar{v} \neq \bar{0}$ and $\epsilon > 0$.