Lecture #17 : 11-13-18

Recall: For data $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$,

we aim to find a hyperplane that simultaneously is nearly margin maximizing and separating.

- Let w, b parametrize a hyperplane as $\{xw^T = b\}$. Then we seek to minimize

$$F(w,b) = \|w\|_2^2 + \lambda \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b)) \quad \text{(hinge loss)}$$

$$G(w,b) = \|w\|_2^2 + \lambda \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b))^2 \quad \text{(quadratic loss)}$$

- λ is a regularization/tuning parameter that balances between the two terms.

Q: How to solve $(w^*, b^*) = \underset{(w,b)}{\operatorname{argmin}} \|w\|_2^2 + \lambda \sum_{i=1}^n \max(0, (1 - y_i(w^T x_i + b)))$?

- We develop a dual formulation to this problem.

- Need to develop a bit of the theory of optimization.
- Recall that if a function $f: \mathbb{R} \to \mathbb{R}$ is smooth, it can be minimized by computing the gradient and setting $= 0$: $\nabla f = 0$. Then the second derivative matrix (Hessian) may be used to classify any zeroes of ∇f.

- Things are more delicate if we add a constraint: $x \in \mathbb{R}^d$, $f: \mathbb{R}^d \to \mathbb{R}$, $g: \mathbb{R}^d \to \mathbb{R}^d$

$$\text{minimize } f(x) \text{ subject to } h(x) \geq 0$$

i.e.

$$(h(x))_i \geq 0 \quad \text{for all } i = 1, \dots, d.$$

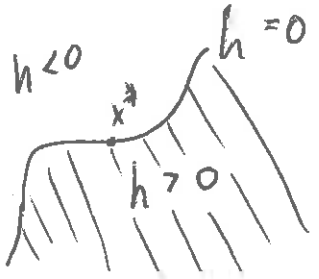
- The constraint $h(x) \geq 0$ defines a feasible set $R = \{x \in \mathbb{R}^d \mid h(x) \geq 0\}$.
- R contains an interior $\mathring{R} = \{x \mid (h(x))_i \overset{\forall i}{>} 0\}$ and a boundary

$$\partial R = \{x \mid (h(x))_i = 0, \text{ some } i\}.$$

- Clearly any interior point $x^* \in \mathring{R}$ is a minimizer of f only if $\nabla f(x^*) = 0$. This is because the derivative is a locally defined function, and if $x^* \in \mathring{R}$, $B(x^*, \varepsilon) \subset \mathring{R}$ for some $\varepsilon > 0$.
- The boundary is much more delicate. We cannot say $\nabla f(x_*) = 0$, since we could just be cutting off f off away from its minimizer, i.e. $f(x) = |x|$ has a minimizer at $x = 1$ under the constraint $x \geq 1$, but $\nabla f = f' \neq 0$ at $x = 1$.

Let's visualize things when $h: \mathbb{R}^2 \to \mathbb{R}$, i.e. we have a single constraint.

Suppose h is smooth:



Recall that if $x^* \in \{x \mid h(x)=0\}$ (or any level set), then $\nabla h(x^*)$ points orthogonal to the tangent direction: Suppose h is smooth, so that the level set gives an implicit relation between the inputs of h: $h(x,y)=0$

$$\Leftrightarrow h(x, y(x))=0,$$

where $y(x)$ is an implicit curve. Then differentiating in x:

$$\nabla_x h(x,y) = 0$$

$$\Leftrightarrow \nabla_x h(x, y(x)) = 0$$

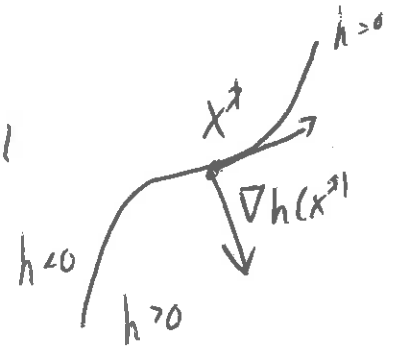
$$\Leftrightarrow \frac{\partial h}{\partial x} + \frac{\partial h}{\partial y} \cdot \frac{\partial y}{\partial x} = 0$$

$$\Leftrightarrow \frac{\partial y}{\partial x} = \frac{-\frac{\partial h}{\partial x}}{\frac{\partial h}{\partial y}}$$, which is the tangent direction of the level set.

On the other hand, $\nabla h = \left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right)$, which has tangent direction (rise over run)

$\frac{\partial h}{\partial y} / \frac{\partial h}{\partial x}$ $\Rightarrow$ tangent direction of the level set is orthogonal to the ∇h.

So,



In particular, if $v, \|v\|_2 = 1$ forms an angle greater than $\frac{\pi}{2}$ in absolute value with $\nabla h(x^*)$, then v points "out" of $\{h \geq 0\}$. Otherwise, it points in.

Recall that $\frac{d}{dt}\left(h(x^* + tv)\right)\Big|_{t=0} = v \cdot \nabla h(x^*)$ is the directional derivative.

- This allows us to formulate a sufficient condition for having a minimum to f at x^*: no increase in directions pointing inside $R = \{h \geq 0\}$:

if $\underbrace{v \cdot \nabla h(x^*) > 0}_{v \text{ points inside } R}$, then $\underbrace{v \cdot \nabla f(x^*) \geq 0}_{f \text{ is non-decreasing in the } v \text{ direction}}$

- In other words, $\nabla h(x^*)$ and $\nabla f(x^*)$ point in the same direction! So, $\exists\, \lambda \geq 0$ s.t. $\nabla h(x^*) = \lambda \nabla f(x^*)$; this is just like Lagrange multipliers from multivariate calculus.

- The condition $\exists\, \lambda \geq 0$ such that $\nabla h(x^*) = \lambda \nabla f(x^*)$ is the Karush-Kuhn-Tucker (KKT) condition. It is a necessary condition for x^* being a local minimizer on the boundary of R.

- This generalizes to the KKT Theorem, which provides a necessary condition for the constrained optimization minimize $f(x)$ subject to $g(x) \geq 0$:

Theorem (KKT): Let x^* be a solution to minimize $f(x)$ subject to $g(x) \geq 0$, such that $(g(x^*))_i = 0$ for some $i \in I \subset \{1, \dots, d\}$. Then $\nabla f(x^*) = \sum_{i \in I} \lambda_i \nabla h(x^*)$, some $\lambda_i \geq 0$.

Remark: There may be multiple solutions to a general optimization problem. Indeed, if R is bounded and $f, h = (h_1, \dots, h_d)$ are continuous. Then the "$\{g(x) \geq 0\}$" minimization problem minimize $f(x)$ s.t. $x \in R$ has a solution.

Problem: Could be non-unique, and there could be many local minima to mess us up.

If f is convex, there is exactly one local minimum, which addresses the above ambiguity.

Defn: Let $R \subset \mathbb{R}^d$. R is convex as a set if $\forall x, y \in R$, $[tx + (1-t)y] \in R$ for all $t \in [0,1]$.

Defn: Let $f: \mathbb{R} \to \mathbb{R}$ be defined on a convex set R. We say f is convex if $f(tx + (1-t)y) \leq f(x) + (1-t)f(y)$, $\forall x, y \in R$, $\forall t \in [0,1]$. We say f is strictly convex if $\leq$ may be replaced by $<$.

Exercise: f is strictly convex on R iff $Hf(x)$ is positive definite for all x, where $Hf(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{i,j=1}^{d}(x)$ is the Hessian matrix at x.

- So, if we can prove a feasible set R is bounded and convex, and that f is convex on R (for example by analyzing the Hessian on R), then the constrained minimization problem minimize $f(x)$ s.t. $x \in R$ has a unique solution.