UMC203: AI and ML

Naman Mishra

January 2024

Contents

1 Convex Optimisation

Lecture 07: Thu 01 Feb '24

Setting:

$$\begin{aligned}
&\mathcal{D} = \left\{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i \in [N]\right\} \\
&\mathcal{D} \sim P^{(N)} \\
&\mathcal{X} \subseteq \mathbb{R}^d \\
&\mathcal{Y} = \{-1, 1\}.
\end{aligned}$$

Suppose there exists a $w^* \in \mathbb{R}^d$ such that for each $i \in [N]$, $\operatorname{sgn}(w^{*\top} x^{(i)}) = y^{(i)}$.
Then the algorithm described in the previous lecture will find such a w^* in at most $\frac{R^2}{\gamma^{*2}}$ iterations, where R is the maximum norm of $x^{(i)}$s and $\gamma^* = \min_{i \in [N]} \frac{|w^{*\top} x^{(i)}|}{\|w^*\|}$.

Let $\mathcal{D}$ be linearly separable and let the Perceptron algorithm return a classifier $h_{\mathcal{D}}^{(p)}$. We have risk $R(h_{\mathcal{D}}^{(p)}) = \Pr(h_{\mathcal{D}}^{(p)}(x) \neq y)$ (under what distribution?). We compute the *expected generalization error* by a classifier returned by the Perceptron algorithm acting on a linearly separable sample of size N drawn iid from P. That is, we compute

$$\mathop{\mathbb{E}}_{\mathcal{D} \sim P^N} [R(h_{\mathcal{D}}^{(p)})].$$

This is hard!

We will instead compute the proxy $\overline{R}_{\mathcal{D}}^{LOO}(A)$, where A is an algorithm acting on a sample $\mathcal{D}$ of size m, returning a classifier $h_{\mathcal{D}}^{A}$.

$\overline{R}_{\mathcal{D}}^{LOO}(A)$ is the *leave-one-out* error of A on $\mathcal{D}$. That is,

$$\overline{R}_{\mathcal{D}}^{\text{LOO}}(A) = \frac{1}{m} \sum_{i=1}^{m} \left[h_{\mathcal{D}_{(i)}}^{A}\big(x^{(i)}\big) \neq y^{(i)} \right],$$

where $\mathcal{D}_{(i)} = \mathcal{D} \setminus \big\{ (x^{(i)}, y^{(i)}) \big\}$.

We want to compute the expected value of $\overline{R}_{\mathcal{D}}^{LOO}(A)$ over all samples $\mathcal{D}$ of size m drawn iid from P.

$$\mathop{\mathbb{E}}_{\mathcal{D} \sim P^m} [\overline{R}_{\mathcal{D}}^{\text{LOO}}(A)] = \frac{1}{m} \mathop{\mathbb{E}}_{\mathcal{D} \sim P^m} \sum_{i=1}^{m} \left[h_{\mathcal{D}_{(i)}}^{A}\big(x^{(i)}\big) \neq y^{(i)} \right]$$

Since the samples are iid, we have

$$\mathop{\mathbb{E}}_{\mathcal{D} \sim P^m} = \mathop{\mathbb{E}}_{\mathcal{D} \sim P^{m-1}} \mathop{\mathbb{E}}_{(x^{(i)}, y^{(i)}) \sim P}$$

We first compute the inner expectation.

$$\begin{aligned} \mathop{\mathbb{E}}_{(x^{(i)}, y^{(i)}) \sim P} \left[h_{\mathcal{D}_{(i)}}^{A}\big(x^{(i)}\big) \neq y^{(i)} \right] &= \Pr \left(h_{\mathcal{D}_{(i)}}^{A}\big(x^{(i)}\big) \neq y^{(i)} \right) \\ &= R(h_{\mathcal{D}_{(i)}}^{A}) \end{aligned}$$

So we have

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathcal{D} \sim P^m} [\overline{R}_{\mathcal{D}}^{\text{LOO}}(A)] &= \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\mathcal{D}_{(i)} \sim P^{m-1}} R(h_{\mathcal{D}_{(i)}}^{A}) \\ &= \mathop{\mathbb{E}}_{\mathcal{D} \sim P^{m-1}} R(h_{\mathcal{D}}^{A}). \end{aligned}$$

Lecture 08: Tue 06 Feb '24

1 Convex Optimisation

Definition 1.1 (Convex function)**.** A set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0, 1]$,

$$(1 - \lambda)x + \lambda y \in C.$$

A function $f \colon C \to \mathbb{R}$ over a convex set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0, 1]$,

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y).$$

Theorem 1.2. Let $f \in C^1(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Then f is convex iff

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x \rangle$$

Notation. Let A and B be symmetric matrices. We write $A \succeq B$ if $A - B$ is positive semidefinite.

Theorem 1.3. $\succeq$ is a partial order.

Proof.

- Reflexivity: $A - A = 0 \succeq 0$.
- Antisymmetry: $A - B \succeq 0$ and $B - A \succeq 0$ implies $A - B = 0$, since λ and $-\lambda$ are both nonnegative for each eigenvalue λ of the difference.
- Transitivity: Suppose $A \succeq B \succeq C$. Then for all u,

$$\begin{aligned} &\langle u,\, (A - B)u \rangle \geq 0 \\ &\langle u,\, (B - C)u \rangle \geq 0 \\ \implies &\langle u,\, (A - C)u \rangle = \langle u,\, (A - B + B - C)u \rangle \\ &\qquad\qquad\qquad = \langle u,\, (A - B)u \rangle + \langle u,\, (B - C)u \rangle \\ &\qquad\qquad\qquad \geq 0. \end{aligned}$$

□

Theorem 1.4. Let $f \in C^2(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Let $H(x) = (\operatorname{Hess} f)(x)$. Then f is convex iff

$$H(x) \succeq 0 \quad \forall\, x \in C.$$

Definition 1.5 (Convex optimisation problem)**.** Let $f\colon \mathbb{R}^d \to \mathbb{R}$ and $f_i\colon \mathbb{R}^d \to \mathbb{R}$ be convex functions for each $1 \leq i \leq m$. Let $(a_j)_{j=1}^n \subseteq \mathbb{R}^d$ and $(b_j)_{j=1}^n \subseteq \mathbb{R}$. The *convex optimisation problem* is to find

$$\min_{x \in \mathbb{R}^d} f(x) \quad \text{such that} \quad \begin{cases} f_i(x) \leq 0 \text{ for all } i \in [m], \\ \langle a_j,\, x \rangle = b_j \text{ for all } j \in [n]. \end{cases}$$

Definition 1.6 (Lagrangian)**.** Let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$. The *Lagrangian* of the convex optimisation problem is

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \mu_j \left(\langle a_j,\, x\rangle - b_j\right).$$

We say that x^* is a *KKT point* if there exist λ and μ such that

$$\begin{aligned} \nabla_x L(x^*, \lambda, \mu) &= 0, \\ \langle a_j,\, x^*\rangle - b_j &\leq 0 \quad \forall\, j \in [n], \\ f_i(x^*) &\leq 0 \quad \forall\, i \in [m], \\ \lambda_i f_i(x^*) &= 0 \quad \forall\, i \in [m]. \end{aligned}$$

Theorem 1.7. If x^* is a KKT point for the convex optimisation problem, then x^* is a minimiser of the problem (for most problems).

Example. Consider the convex optimisation problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|x - z\|^2 \quad \text{such that} \quad \langle w,\, x\rangle + b = 0.$$

The Lagrangian is

$$L(x, \mu) = \frac{1}{2}\|x - z\|^2 + \mu(\langle w,\, x\rangle + b).$$

The KKT conditions are

$$\begin{aligned} \nabla_x L(x^*, \mu) = x - z + \mu w &= 0, \\ \implies x^* &= z - \mu w, \\ \langle w,\, x^*\rangle + b &= 0 \\ \implies \langle w,\, z - \mu w\rangle + b &= 0 \\ \implies \langle w,\, z\rangle - \mu\|w\|^2 + b &= 0 \end{aligned}$$

So the minimiser is

$$x^* = z - \frac{(\langle w,\, z\rangle + b)}{\|w\|^2} w.$$

This is the orthogonal projection of z onto the hyperplane.

Lecture 09: Thu 08 Feb '24