Lec. 18 (11/22/2022)

MAP estimation contd. from last class :

Shadi covered in last class :

$$p_{X|Y}(\underline{x}, \underline{y}) = \frac{p(\underline{x}, \underline{y})}{p_Y(\underline{y})} = \frac{p_{Y|X}(\underline{x}, \underline{y})\, p_X(\underline{x})}{p_Y(\underline{y})}$$

Now, $\hat{\underline{x}}^*_{MAP} = \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}}\; p_{X|Y}(\underline{x}, \underline{y})$

$$= \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}} \frac{p_{Y|X}(\underline{x}, \underline{y})\, p_X(\underline{x})}{\cancel{p_Y(\underline{y})}} \rightarrow \text{indep. of } \underline{x}$$

(so just drop it)

$$= \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}} \left\{ \underbrace{\log p_{Y|X}(\underline{x}, \underline{y}) + \log p_X(\underline{x})}_{\log \text{ of Joint } p(\underline{x}, \underline{y})} \right\}$$

Taking $\log(\cdot)$ of the objective

Example: Linear measurements with i.i.d. noise

$$y_i = \underline{a}_i^T \underline{x} + v_i, \quad i = 1, \dots, m,$$

i.i.d. noise with PDF $p_v(\cdot)$ on $\mathbb{R}$

Now we also have:

$$\underline{x} \sim \underbrace{p_X(\underline{x}) \text{ on } \mathbb{R}^n}_{\text{prior PDF}}$$

$\therefore$ Joint PDF $p(\underline{x}, \underline{y}) = p_X(\underline{x}) \prod_{i=1}^{m} p_v(y_i - a_i^T \underline{x})$

$$\therefore \hat{x}^*_{MAP} = \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}} \left\{ \underbrace{\log p_x(\underline{x})}_{\text{Extra term compared to MLE}} + \sum_{i=1}^{m} \log p_v(y_i - a_i^T \underline{x}) \right\}$$

In particular, suppose $v_i \sim \underbrace{\text{Unif}[-a, +a]}_{p_v(\cdot)} = \frac{1}{2a}$

$$\underline{x} \sim \underbrace{\mathcal{N}(\underline{\mu}, P)}_{p_x(\cdot)}$$

$\underline{\mu}$: mean vector $\in \mathbb{R}^n$

P: covariance matrix $\in S^n_{++}$

$$p_x(\cdot) = \frac{1}{\sqrt{(2\pi)^n \det(P)}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})^T P^{-1} (\underline{x}-\underline{\mu})\right)$$

We need :

$$-a \leq y_i - \underline{a}_i^T \underline{x} \leq +a \quad \forall i = 1, \ldots, m$$

$$\iff \quad \| A \underline{x} - \underline{y} \|_\infty \leq a$$

where A has rows $\underline{a}_i^T$.

$$\therefore \quad \hat{x}^*_{MAP} = \underset{\underline{x} \in \mathbb{R}^n}{\operatorname{argmax}} \left\{ -\frac{1}{2} (\underline{x} - \underline{\mu})^T P^{-1} (\underline{x} - \underline{\mu}) + \text{constant} \right\}$$

$$\text{s.t.} \quad \| A \underline{x} - \underline{y} \|_\infty \leq a$$

next pg.

$$\Leftrightarrow \quad \hat{\underline{x}}^*_{MAP} = \underbrace{\begin{aligned} &\underset{\underline{x}\in\mathbb{R}^n}{\arg\min} \ \frac{1}{2}(\underline{x}-\underline{\mu})^T P^{-1}(\underline{x}-\underline{\mu}) \\ &\text{s.t.} \quad \|A\underline{x}-\underline{y}\|_\infty \le a \end{aligned}}_{QP} .$$

Classification / Discrimination problems:

Given 2 sets of datapoints in $\mathbb{R}^n$

$\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$

and $\{\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_M\}$

Mathematically, we want to find a function $f: \mathbb{R}^n \mapsto \mathbb{R}$ within certain class of functions s.t.

$$f(\underline{x}_i) > 0, \quad i = 1, \ldots, N$$

$$f(\underline{y}_i) < 0, \quad i = 1, \ldots, M$$

Then, the zero level set of f, i.e.,

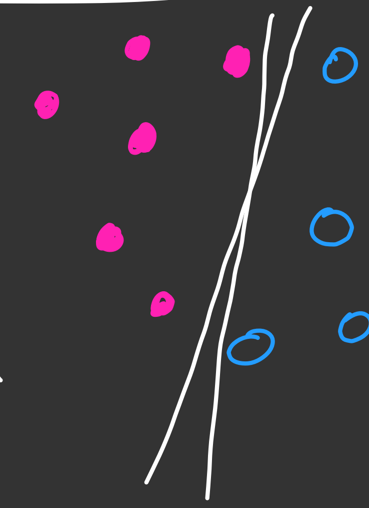
$\{ \underline{x} \mid f(\underline{x}) = 0 \}$ discriminates/classifies/separates the two sets of data.

Simplest scenario: Linear classifier/discriminator:

$$f(\underline{x}) = \langle \underline{a}, \underline{x} \rangle - b$$

i.e., $\underline{a}^T \underline{x}_i - b > 0 \ \forall \ i = 1, \dots, N$

and $\underline{a}^T \underline{y}_i - b < 0 \ \forall \ i = 1, \dots, M$



When is a problem linearly classifiable?

Theorem of alternatives for linear inequalities (text section: 5.8.3)

NOT solvable if and only if

$$\exists\ \underline{\lambda}, \underline{\tilde{\lambda}} \text{ s.t. } \underline{\lambda} \succcurlyeq \underline{0},\ \underline{\tilde{\lambda}} \succcurlyeq \underline{0},\ \mathbb{1}^T \underline{\lambda} = 1,$$

$$\mathbb{1}^T \underline{\tilde{\lambda}} = 1,$$

$$\sum_{i=1}^{N} \lambda_i \underline{x}_i = \sum_{i=1}^{M} \tilde{\lambda}_i \underline{y}_i$$

$\iff$ (Geometric interpretation)

$\exists$ a point in $\text{conv}\{\underline{x}_1, \dots, \underline{x}_N\}$ AND $\text{conv}\{\underline{y}_1, \dots, \underline{y}_M\}$

$\iff$ linearly classifiable/separable if and only if the two convex hulls do NOT intersect.

Robust linear classifier/discriminator (RLD)

Find $\underline{\underline{\text{optimal}}}$ $(\underline{a}, b)$ in $f(\underline{x}) = \langle \underline{a}, \underline{x} \rangle - b$

one that maximizes the gap between

> 0 values @ $\underline{x}_i$, and < 0 values @ $\underline{y}_i$

$$\Leftrightarrow \quad \underset{t,\, \underline{a},\, b}{\text{maximize}} \quad t$$

$$\text{s.t.} \quad \begin{aligned} &\langle \underline{a}, \underline{x}_i \rangle - b \geq t, \quad i = 1, \dots, N \\ &\langle \underline{a}, \underline{y}_i \rangle - b < t, \quad i = 1, \dots, M \\ &\|\underline{a}\|_2 \leq 1 \end{aligned}$$

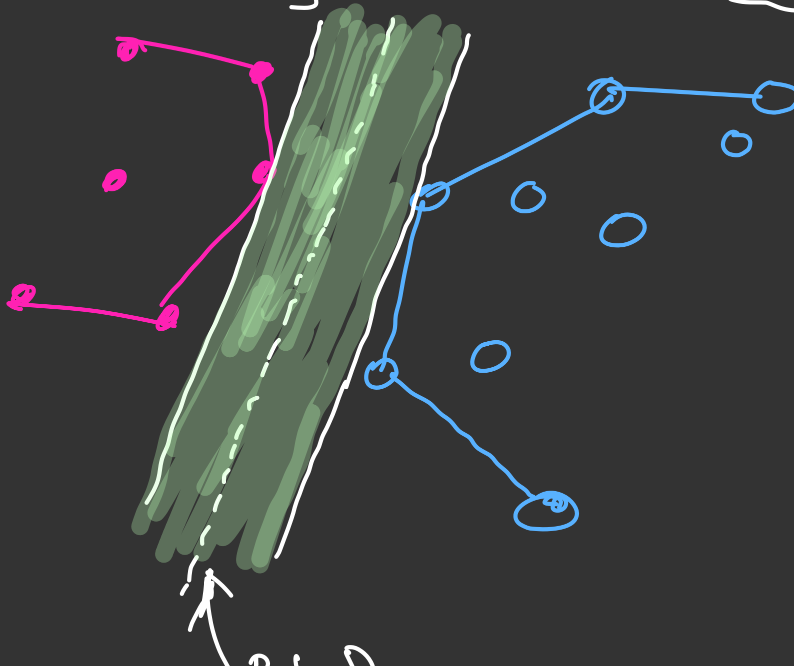
So the maximizer/argmax $(t^*, \underline{a}^*, b^*)$ define the optimal hyperplane/linear classifier

- linearly classifiable $\Longleftrightarrow$ $t^* > 0$
- We can prove that $\|\underline{a}^*\|_2 = 1$.
- If $\|\underline{a}\|_2 = 1$ then $\langle \underline{a}, \underline{x}_i \rangle - b$ is the Euclidean dist. from $\underline{x}_i$ to the separating hyp. plane $\mathcal{H} := \{ \underline{z} \in \mathbb{R}^n \mid \langle \underline{a}, \underline{z} \rangle = b \}$
- Similarly, $b - \langle \underline{a}, \underline{y}_i \rangle$ is the dist. from $\underline{y}_i$ to $\mathcal{H}$.

$\therefore$ RLD finds the maximal separator

$\Updownarrow$

Computing the "thickest slab" between two given datasets



RLD solution

(optimal hyperplane)

$$t^* = \frac{1}{2} \times \text{distance between the two conv. hulls}$$

(see text: p. 424-425)

What to do if we know that the 2 datasets are NOT linearly classifiable?

(Can be checked via Thm. of Alternatives)

Idea # 1

Approximate linear classifier

$\Updownarrow$

Minimize the # of misclassification errors

Idea # 2

Exact nonlinear classifier

Example: polynomial $f(\underline{x})$ s.t.

$$f(\underline{x}_i) > 0 \quad \forall\, i = 1, \dots, N$$
$$f(\underline{y}_i) < 0 \quad \forall\, i = 1, \dots, M$$

Idea #1

$\Downarrow$

Support vector machine (SVM)

Recall: linear classifier:

$$\langle \underline{a}, \underline{x}_i \rangle - b > 0 \ \forall i = 1, \dots, N$$

$$\langle \underline{a}, \underline{y}_i \rangle - b < 0 \ \forall i = 1, \dots, M$$

$$\Longleftrightarrow \langle \underline{a}, \underline{x}_i \rangle - b \geqslant 1 \ \forall i = 1, \dots, N$$

$$\langle \underline{a}, \underline{y}_i \rangle - b \leqslant -1 \ \forall i = 1, \dots, M$$

Relax this conditions for allowing misclassification

Idea #2

If $f(\cdot)$ is a polynomial on $\mathbb{R}^n$ with degree $\leq d$,

i.e.,

$$f(\underline{x}) = \sum_{i_1 + i_2 + \dots + i_n \leq d} a_{i_1 i_2 \dots i_d} x_1^{i_1} \dots x_n^{i_n}$$

its zero level set

$$\{ \underline{x} \in \mathbb{R}^n \mid f(\underline{x}) = 0 \}$$

is an algebraic surface

Relaxation for approx. lin. classification:

Introduce: $u_1, \ldots, u_N \geq 0$
and $v_1, \ldots, v_M \geq 0$ $\Big\}$

such that $\langle \underline{a}, \underline{x}_i \rangle - b \geq 1 - u_i, \quad i = 1, \ldots, N$

$\langle \underline{a}, \underline{y}_i \rangle - b \leq -(1 - v_i), \quad i = 1, \ldots, M$

Original/exact linear classification:

$$\underline{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} = \underline{0}$$

$$\underline{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_M \end{pmatrix} = \underline{0}$$

$\therefore$ $\underline{u}$ and $\underline{v}$ measure how much the inequalities are violated.

Problem:

$$\underset{\substack{\underline{u} \in \mathbb{R}^N \\ \underline{v} \in \mathbb{R}^M \\ \underline{a} \in \mathbb{R}^n \\ b \in \mathbb{R}}}{\text{minimize}} \ \underbrace{\|\underline{u}\|_{\ell_0(\mathbb{R}^N)}}_{\text{cardinality/number of nonzero entries in vector } \underline{u}} + \underbrace{\|\underline{v}\|_{\ell_0(\mathbb{R}^M)}}_{\text{likewise for } \underline{v}}$$

nonconvex hard combinatorial problem

$$\text{s.t.} \quad \langle \underline{a}, \underline{x}_i \rangle - b \geq 1 - u_i \ \ \forall \, i = 1, \ldots, N$$

$$\langle \underline{a}, \underline{y}_i \rangle - b \leq -(1 - v_i) \ \ \forall \, i = 1, \ldots, M$$

$$\underline{u} \geq \underline{0}$$

$$\underline{v} \geq \underline{0}$$

<u>Heuristics</u>: one of the heuristics is to convexify the objective: replace the ℓ_0 norms by ℓ_1 norms.

then the prev. page objective becomes: $\underbrace{\mathbb{1}^T \underline{u} + \mathbb{1}^T \underline{v}}_{\text{LP}}$