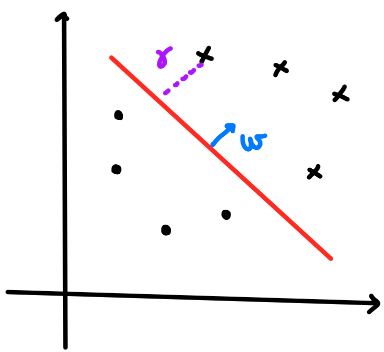
SUPPORT VECTOR MACHINES

7/10/2023

IS A CLASSIFICATION ALGORTHM

- $y \in \{-1, +1\}$
- $\Theta = (\underbrace{\Theta_0}_{b \in \mathbb{R}}, \underbrace{\Theta_1, \dots, \Theta_d}_{\vec{w} \in \mathbb{R}^d}) \in \mathbb{R}^{d+1}$

SUPPOSE WE HAVE A DATASET



HYPOTHESIS CLASS

$h_{w,b}(x) = \text{sign}(w^T x + b)$

HYPERPLANE

$r: w^T x + b = 0$, $w \perp r$ $\longrightarrow \begin{cases} (w^T x^{(i)} + b) \geq 0 & y = 1 \quad \bullet \\ (w^T x^{(i)} + b) < 0 & y = -1 \quad \times \end{cases}$

MARGIN

WE CAN SEE IT AS THE DISTANCE BETWEEN r AND A DATA POINT. WE WANT THE LARGEST VALUE POSSIBLE OF THE MARGIN.

→ FUNCTIONAL $\quad \hat{\gamma}^{(i)} = y^{(i)}\left[w^T x^{(i)} + b\right] \to \hat{\gamma} = \min_{i=1,\dots,n} \hat{\gamma}^{(i)}$

PROBLEM, IF I WANT TO SCALE w AND b BY t, THEN $\hat{\gamma}^{(i)}$ GETS t-UPLED (NOT GOOD) (WHILE $h_{w,b}(x)$ REMAINS THE SAME). WE NEED THE GEOMETRIC

→ GEOMETRIC $\quad \gamma^{(i)} = y^{(i)}\left[\frac{w^T x^{(i)}}{\|w\|} + \frac{b}{\|w\|}\right] \to \gamma = \min_{i=1,\dots,n} \gamma^{(i)}$ (IF t-SCALES, GETS t-CANCELED)

IN ORDER TO FIND THE OPTIMAL CLASSIFIER WE NEED TO APPLY A CONVEX OPTIMIZATION PROBLEM

OPERATION OF OP.

OBJECTIVE

VARIABLES W.R.T. OPTIMIZE

$$\max_{\gamma, w, b} \quad \gamma$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \dots, n$$

$$\|w\| = 1$$

CONSTRAINTS

$$\Downarrow$$

PRIMAL OPTIMIZATION PROBLEM

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

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

$$\Downarrow$$

DUAL OPTIMIZATION PROBLEM

W/ LAGRANGE MULTIPLIERS

$$\max_{\alpha} \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

SOLVING IT

α - OUTPUT FROM SOLUTION

WE CAN NOW MAKE A PREDICTION

$$(w^T x + b) = \left(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}\right)^T x + b = \sum_{i=1}^{n} \alpha_i y^{(i)} \underbrace{\langle x^{(i)}, x \rangle}_{\text{KERNELIZED}} + b$$

$K(x^{(i)}, x^{(j)})$

SUPPORT VECTORS α'S

α ARE MOSTLY $\vec{0}$, THE MORE THEY ARE DEFINED, THE MORE THE IMPACT ON THE LINEAR CLASSIFIER (SMALL MARGIN)

BAYESIAN METHODS

- SO FAR WE WORKED W/ FREQUENTIST APPROACH → MLE, θ IS CONSTANT VALUED AND UNKNOWN
- BAYESIAN APPROACH → $\hat{\theta} = \arg\max_{\theta} \ell(\theta, S)$
  → $S = \{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ (TRAINING SET)

→ θ IS A RANDOM VARIABLE AND UNKNOWN

① DEFINE A PRIOR DISTRIBUTION $P(\theta)$ EVEN BEFORE LOOKING AT DATA
→ ② OBSERVE THE DATA S
→ ③ DEFINE A POSTERIOR DISTRIBUTION $P(\theta|S)$ THAT UPDATES BELIEFS ON θ BASED ON THE DATA
→ ④ MAKE PREDICTIONS $h_\theta(x) = g(\theta^T x) = \mathbb{E}[y|x]$

$$P(\theta|S) = \frac{P(S|\theta)\,P(\theta)}{P(S)}$$

POSTERIOR PREDICTIVE DISTRIBUTION (DISTRIBUTION, NOT PREDICTION)

$$P(y_*|x_*, S) = \underbrace{\int_\theta}_{\text{AVERAGING}} P(y_*|x_*, \theta)\, \overbrace{P(\theta|S)}^{\text{POSTERIOR DISTRIBUTION}}\, d\theta$$

WITH y_*, x_* BEING UNKNOWN

$$= \mathbb{E}_{\theta \sim P(\theta|S)}\left[P(y_*|x_*, \theta)\right]$$

WE CONSIDER EVERY VALUE OF θ GIVEN FROM THE DISTRIBUTIONS AS LINEAR COMBO OF THE ONES GIVEN IN THE POSTERIOR DISTRIBUT.

$$P(y_* | x_*, S) = \int_\theta P(y_*, \theta | x_*, S)\, d\theta$$

LET'S NOW APPLY THE CHAIN RULE

$$= \int_\theta \underline{P(\theta | x_*, S)}\ \underline{P(y_* | \theta, x_*, S)} = \int_\theta \underline{P(\theta | S)}\ \underline{P(y_* | \theta, x_*)}\, d\theta = \int_\theta \underline{P(y_* | x_*, \theta)}\ \underline{P(\theta | S)}\, d\theta$$

THE UNKNOWN θ IS INDIPENDENT FROM TEST EXAMPLES X

PROBLEMS

1. HOW DO WE CHOOSE A PRIOR?
2. HOW DO WE GO FROM PRIOR TO POSTERIOR?

$P(\theta)$ — PRIOR DISTRIBUTION

$P(S | \theta)$ — LIKELIHOOD (THIS DISTRIBUTION DEFINES OUR MODEL)

WE CAN GET TO THE POSTERIOR THANKS TO BAYES

$$P(\theta | S) = \frac{P(\theta)\, P(S | \theta)}{P(S)}$$

$$\boxed{P(\theta | S) = \frac{P(\theta)\, P(S | \theta)}{\int_\theta P(\theta) P(S | \theta)\, d\theta}}$$

EX MONTE CARLO APPROACH

$\theta^{(0)}, \theta^{(1)}, \dots, \theta^{(k)}$

GIBBS SAMPLING : EXTRACTION OF FEATURES

$$\frac{1}{N} \sum_{i=1}^{n} P(y_* | x_*, \theta^{(i)}) \approx \mathbb{E}_{\theta \sim P(\theta | S)} [P(y_* | x_*, \theta)]$$

USUALLY WE DON'T HAVE CLOSED-FORM EXPRESSIONS, BUT APPROXIMATIONS OF THE POSTERIOR DISTR. FUNCTION