Amrita Sundari V
amrita95@tamu.edu
16 Nov 2017.

WRITTEN ASSIGNMENT-3

1) Show that the hard-SVM rule namely,

$$\underset{(w,b):\|w\|=1}{\operatorname{argmax}} \ \min_{i\in[M]} |\langle w, x_i\rangle + b| \quad s.t\ \forall i,\ y_i(\langle w, x_i\rangle + b) > 0.$$

is equivalent to the following formulation

$$\underset{(w,b):\|w\|=1}{\operatorname{argmax}} \ \min_{i\in[m]} y_i(\langle w, x_i\rangle + b).$$

Let G be the set of all ~~hyperplanes~~ half spaces (w, b) such that $y_i(\langle w, x_i\rangle + b) > 0\ \forall i$

i.e $G = \{(w,b) : \forall i\ y_i(\langle w, x_i\rangle + b) > 0\}$

If there exist a half space (w, b) then $y_i(\langle w, x_i\rangle + b) > 0\ \forall i \in [m]$ for the training set given S.

$y_i \in \{+1, -1\}$

$|\langle w, x_i\rangle + b|$ is equivalent to $y_i(\langle w, x_i\rangle + b)$ since $y_i(\langle w, x_i\rangle + b)$ is always positive ~~given~~ (condition for Hard SVM) and therefore it will replace the ~~absolute~~ modulus function (which also returns the +ve value of the argument)

$$\therefore \underset{(w,b):\|w\|=1}{\operatorname{argmax}} \ \min_{i\in[m]} |\langle w, x_i\rangle + b| \equiv \underset{(w,b):\|w\|=1}{\operatorname{argmax}} \ \min_{i\in[m]} y_i(\langle w,$$

$$s.t\ y_i(\langle w, x_i\rangle + b)$$

Scanned by CamScanner

2. Weak Duality: Prove that for any function f of 2 vector variables $x \in X$, $y \in Y$ it holds that

$$\min_{x \in X} \max_{y \in Y} f(x,y) \geq \max_{y \in Y} \min_{x \in X} f(x,y)$$

Proof:

Let us consider $y^* = \underset{y \in Y}{\operatorname{argmax}} f(x,y)$.

Then we know that

$$f(x,y) \leq f(x,y^*)$$

$$\Rightarrow \min_{x \in X} f(x,y) \leq \min_{x \in X} f(x,y^*)$$

$$\Rightarrow \max_{y \in Y} \min_{x \in X} f(x,y) \leq \max_{y \in Y} \min_{x \in X} f(x,y^*)$$

$$= \min_{x \in X} f(x,y^*)$$

But $f(x,y^*) = \max_{y \in Y} f(x,y)$.

$$\Rightarrow \max_{y \in Y} \min_{x \in X} f(x,y) \leq \min_{x \in X} \max_{y \in Y} f(x,y).$$

Scanned by CamScanner

3. A 2x2 probability table, $p(x_1 = i, x_2 = j) = \theta_{ij}$ with $0 \le \theta_{ij} \le 1$, $\sum_{i=1}^{2} \sum_{j=1}^{2} \theta_{ij} = 1$ is learned using maximal marginal likelihood in which x_2 is never observed. Show that if

$$\theta^{(1)} = \begin{pmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{pmatrix}$$ is given as maximal marginal likelihood solution, then $\theta^{(2)} = \begin{pmatrix} 0.2 & 0.4 \\ 0.4 & 0 \end{pmatrix}$ has the same marginal likelihood score.

with respect to $\theta^{(1)}$. $p(x_1 = 1) = \sum_{j=1}^{2} p(x_1 = 1, x_2 = j)$

$$= p(x_1 = 1, x_2 = 1) + p(x_1 = 1, x_2 = 2)$$

$$= 0.3 + 0.3$$

$$\boxed{p(x_1 = 1) = 0.6}$$

$$p(x_1 = 2) = \sum_{j=1}^{2} p(x_1 = 2, x_2 = j)$$

$$= p(x_1 = 2, x_2 = 1) + p(x_1 = 2, x_2 = 2)$$

$$\boxed{p(x_1 = 2) = 0.2 + 0.2 = 0.4}$$

Similarly w.r.t $\theta^{(2)}$

$$p(x_1 = 1) = \sum_{j=1}^{2} p(x_1 = 1, x_2 = j) = 0.2 + 0.4 = 0.6$$

$$p(x_1 = 2) = \sum_{j=1}^{2} p(x_1 = 2, x_2 = j) = 0.4 + 0 = 0.4.$$

$\therefore$ $\theta^{(1)}, \theta^{(2)}$ has the same marginal likelihood score

4. Consider a mixture of factorised models for vector observations v

$$p(v) = \sum_h p(h) \prod_i p(v_i / h)$$

For assumed i.i.d data v^n, $n = 1, \ldots N$, some observation components may be missed so that, for example the third component of the fixed datapoint v_3^5 is unknown. Show that maximum likelihood training on the observed data corresponds to ignoring components v_i^n that are missing.

Let us consider that there are only 2 components $v = [v_1, v_2]$ and 2 hidden states, $dom(h) = \{h_1, h_2\}$

Let v_2 - missing component.

$$p(v) = \sum_h p(h) . \prod_i p(v_i / h)$$

$$= p(h_1) . p(v_1 / h_1) . p(v_2 / h_1)$$

$$+ p(h_2) . p(v_1 / h_2) . p(v_2 / h_2)$$

Since v_2 is the component missing we need to sum over all the states of v_2, i.e. consider v_2 can take 2 states $\{0, 1\}$

$$= p(h_1) . p(v_1 / h_1) . p(v_2 = 0 / h_1) + p(h_1) . p(v_1 / h_1) . p(v_2 = 1 / h_1)$$

$$+ p(h_2) . p(v_1 / h_1) \; p(v_2 = 0 / h_2) + p(h_2) \, p(v_1 / h_1) . p(v_2 = 1 / h_2)$$

$$\sum_{v_2 = \{0,1\}} p(v_2 / h_2) = 1 \text{ and } \sum_{v_2 = \{0,1\}} p(v_2 / h_1) = 1$$

we get

$$= p(h_1) . p(v_1 / h_1) + p(h_2) . p(v_1 / h_2)$$

Therefore it doesn't affect $p(v)$ which can be written just with known components v_i alone. Therefore it will not affect the maximum likelihood formulation also.

This concept can be extended to N no. of datapts $v^1, v^2, v^3 .. v^n$ and any component v_i^N of those datapoints

5. Consider the term

$$\sum_{n=1}^{N} \langle \log p(h) \rangle_{p^{old}(h/v^n)}$$

we wish to optimise the above with respect to distribut $p(h)$. This can be achieved by defining the lagrangian

$$L = \sum_{n=1}^{N} \langle \log p(h) \rangle_{p^{old}(h/v^n)} + \lambda \left(1 - \sum_h p(h)\right)$$

By differentiating the langrangian with respect to $p(h)$ and using the normalisation constraint $\sum_h p(h) = 1$ show that, optimally.

$$L = \sum_{n=1}^{N} \sum_h \log(p(h)).p^{old}(h/v^n) + \lambda \left(1 - \sum_h p(h)\right)$$

$$\frac{\partial L}{\partial p(h_i)} = \frac{1}{p(h_i)} \sum_{n=1}^{N} p^{old}(h/v^n) - \lambda = 0 \quad \text{—— (1)}$$

$$\lambda . p(h_i) = \sum_{n=1}^{N} p^{old}(h/v^n)$$

$$\sum_h \lambda . p(h_i) = \sum_h \sum_{n=1}^{N} p^{old}(h/v^n)$$

$$\lambda = \sum_{n=1}^{N} \sum_h . p^{old}(h/v^n)$$

$$= \sum_{n=1}^{N} 1$$

$$= N.$$

$$\Rightarrow \lambda = N.$$

Substitute $\lambda = N$ in (1).

$$\cancel{\lambda} = \frac{1}{\cancel{\lambda}} \quad N = \frac{1}{p(h_i)} \sum_{h} p^{old}(h/v^n)$$

$$\boxed{p(h) = \frac{1}{N} \sum_{h} p^{old}(h/v^n)}$$