Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

QUESTION 1

Assignment Number: 2
Student Name: Abhisek Panda
Roll Number: 150026
Date: October 10, 2017

---

subPart 1:

Yes. Observe
entry 4(C. Binns, medium, no, heavy, 0-1, NO) and
entry 6(S. Snape, medium, no, heavy, 0-1, YES)
In these data points, the decisions are different even though all parameters except name of professor are identical. Hence, it must be the case that name of professor was the criteria to distinguish both data points.

subPart 2:

Yes, if there exists a novel algorithm to extract information from the names of the professors then its possible. Otherwise not. On a side note,

subPart 3:

$$\begin{aligned} Entropy(S) &= -\frac{5}{15}\log\frac{5}{15} - \frac{10}{15}\log\frac{10}{15} \\ &= 0.9183 \end{aligned}$$

Information Gain for each attribute is calculated as:

$$Gain(S, A) = Entropy(S) - \Sigma_v \frac{|S_v|}{|S|} Entropy(S_v)$$

where S_v is the subset of S where attribute A gets value v.

Therefore,

$$\begin{aligned} Gain(S, size) &= 0.9183 - \frac{6}{15}0.6500 - \frac{5}{15}0.9710 - \frac{4}{15}1 \\ &= 0.0680 \\ Gain(S, like) &= 0.9183 - \frac{4}{15}1 - \frac{11}{15}0.8454 \\ &= 0.0317 \\ Gain(S, workload) &= 0.9183 - \frac{4}{15}0.8113 - \frac{6}{15}1.0 - \frac{5}{15}0.7219 \\ &= 0.0613 \\ Gain(S, \#meetings) &= 0.9183 - \frac{10}{15}0.9710 - \frac{3}{15}0 - \frac{2}{15}0 \\ &= 0.2710 \end{aligned}$$

Hence, the decision tree should have Number of Meetings as its root since it gives us the highest information gain. For further splitting,

$$\begin{aligned} Gain(S_{0-1}, workload) &= 0.9710 - \frac{2}{10}1 - \frac{3}{10}0 - \frac{5}{10}0.7219 \\ &= 0.4100 \\ Gain(S_{0-1}, research\ area) &= 0.9710 - \frac{3}{10}0.9183 - \frac{7}{10}0.9852 \\ &= 0.0587 \\ Gain(S_{0-1}, size) &= 0.9710 - \frac{1}{10}0 - \frac{5}{10}0.9710\frac{4}{10}1 \\ &= 0.085 \end{aligned}$$

So the next node is workload. After this level, nodes can be split on the basis of majority.

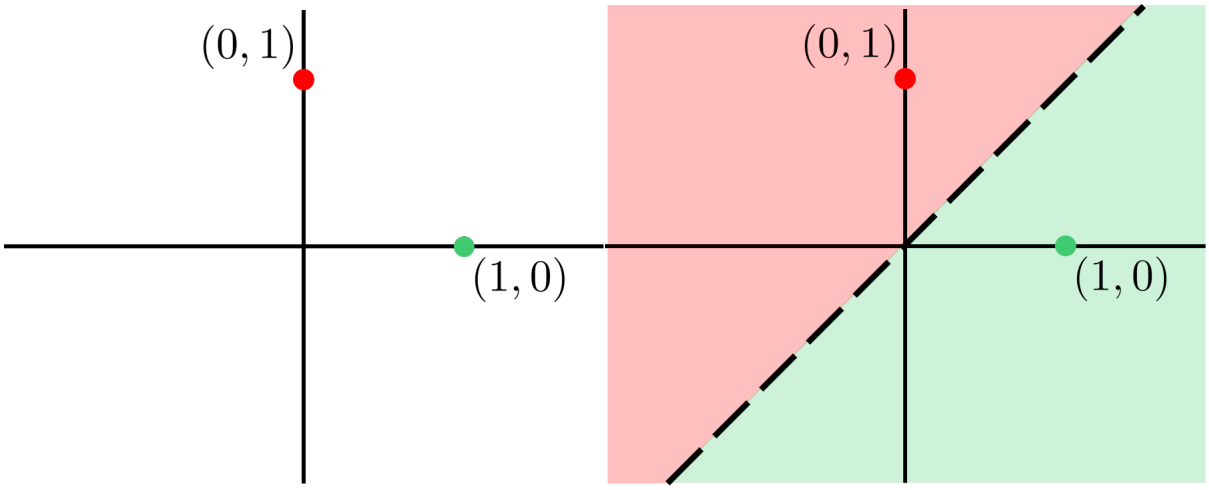

Figure 1: Learning with Prototypes: the figure on the left shows the two prototypes. The figure on the right shows what the decision boundary if the distance measure used is $d(\mathbf{z}^1, \mathbf{z}^2) = \left\|\mathbf{z}^1 - \mathbf{z}^2\right\|_2$, for any two points $\mathbf{z}^1, \mathbf{z}^2 \in \mathbb{R}^2$. The decision boundary in this case is the line $y = x$.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 2
Student Name: Abhisek Panda
Roll Number: 150026
Date: October 10, 2017

QUESTION 2

subPart1:

The products purchased by a customer is represented by a L dimensional vector $\mathbf{S}$ where $\mathbf{S}_j = 1$ implies customer purchased that product. Otherwise, it is 0. To model $\mathbb{P}[b \,|\, \mathbf{x}, \mathbf{\Theta}]$ where b is the bill of the items purchased by the customer $\mathbf{x}$.

$$\mathbb{P}[b \,|\, \mathbf{x}, \mathbf{\Theta}] = \sum_{i=1}^{2^L} \mathbb{P}\left[b \,|\, \mathbf{x}, \mathbf{S}^i, \mathbf{\Theta}\right] \mathbb{P}\left[\mathbf{S}^i \,|\, \mathbf{x}, \mathbf{\Theta}\right]$$

Since the bills have an added gaussian noise, the conditional probability

$$\mathbb{P}\left[b \,|\, \mathbf{x}, S^i, \mathbf{\Theta}\right] = \mathcal{N}(\mu_i, \sigma^2)$$

Also, assuming the choice between various products as independent when conditioned with the customers, $\mathbb{P}\left[\mathbf{S}^i \,|\, \mathbf{x}, \mathbf{\Theta}\right] = \prod_{j=1}^{L} \mathbb{P}\left[\mathbf{S}^i_j \,|\, \mathbf{x}, \mathbf{\Theta}\right]$.
Assume $\mathbb{P}\left[\mathbf{S}^i_j = 1 \,|\, \mathbf{x}, \mathbf{\Theta}\right] = \sigma(w_j^T \mathbf{x}^i)$ where σ is the sigmoid function.
The model for a single customer is given as:

$$P[b^i \,|\, \mathbf{x}^i, \mathbf{\Theta}] = \sum_{j=1}^{2^L} \mathbb{P}\left[b^i \,|\, \mathbf{x}^i, \mathbf{S}^j, \mathbf{\Theta}\right] * \left[\prod_{k=1}^{L} \mathbb{P}\left[\mathbf{S}^j_k \,|\, \mathbf{x}^i, \mathbf{\Theta}\right]\right]$$

Thus, the latent variables are the purchased items vectors, S^is. The model $\mathbf{\Theta}$ consists of weight vectors w_j for each product($j \in [L]$), μ_i for each customer and a σ which is same noise for every customer.

subPart 2:

The likelihood expression for the above objective function is obtained by multiplying the probabilities for each customer:

$$\begin{aligned}
P[\mathbf{b} \,|\, \mathbf{X}, \mathbf{\Theta}] &= \prod_{i=1}^{n} P[b^i \,|\, \mathbf{x}^i, \mathbf{\Theta}] \\
P[\mathbf{b} \,|\, \mathbf{X}, \mathbf{\Theta}] &= \prod_{i=1}^{n} \sum_{j=1}^{2^L} \mathbb{P}\left[b^i \,|\, \mathbf{x}^i, S^j, \mathbf{\Theta}\right] * \left[\prod_{k=1}^{L} \mathbb{P}\left[S^j_k \,|\, \mathbf{x}^i, \mathbf{\Theta}\right]\right] \\
&= \prod_{i=1}^{n} \sum_{j=1}^{2^L} \mathcal{N}\left(\mu_j, \sigma^2\right) * \left[\prod_{k=1}^{L} \mathbb{P}\left[S^j_k \,|\, \mathbf{x}^i, \mathbf{\Theta}\right]\right] \\
\implies \mathbf{\Theta}_{MLE} &= \arg\max_{\mathbf{\Theta}} P[\mathbf{b} \,|\, \mathbf{X}, \mathbf{\Theta}]
\end{aligned}$$

subPart 3:

Hard Assignment Alternating Optimization:

1. Initialize $\mathbf{\Theta}^0$.
2. For $i \in [n]$, update S^i. This can be done by choosing S^i such that the probability $P[S \,|\, \mathbf{x}^i, b^i, \mathbf{\Theta}^t]$ is maximum.

$$S^i = \arg\max_S \; P[S \mid \; \mathbf{x}^i, b^i, \mathbf{\Theta}]$$

where

$$\begin{aligned}
\mathbb{P}\left[S \mid \; \mathbf{x}^i, b^i, \mathbf{\Theta}\right] &\propto P[S \,|\, \mathbf{x}^i, \mathbf{\Theta}] \; P[b^i \,|\, x^i, S, \mathbf{\Theta}] \\
&= \mathcal{N}(\mu^i, \Sigma) \prod_{j=1}^{L} \mathbb{P}\left[S_j \mid \; \mathbf{x}^i, b^i, \mathbf{\Theta}\right] \\
&= \mathcal{N}(\mu^i, \Sigma) \prod_{j=1}^{L} \sigma(\mathbf{w}_j^T \mathbf{x}^i)^{S_j} * [1 - \sigma(\mathbf{w}_j^T \mathbf{x}^i)]^{1-S_j}
\end{aligned}$$

Hence, this reduces to setting all the S_j^i as 1 or 0 depending on whose probabilty is greater than 0.5

3. Update $\mathbf{\Theta}^{t+1}$ using the new obtained S^i.

$$\begin{aligned}
\mathbf{\Theta}^{t+1} &= \arg\max_{\mathbf{\Theta}} P[B, S|X, \mathbf{\Theta}] \\
&= \arg\max_{\mathbf{\Theta}} \prod_{i=1}^{n} P[b^i \,|\, S^i, x^i, \mathbf{\Theta}] \; P[S^i \,|\, x^i, \mathbf{\Theta}] \\
&= \arg\max_{\mathbf{\Theta}} \prod_{i=1}^{n} \mathcal{N}(\mu^i, \sigma^2) \prod_{j=1}^{L} P[S_j^i \,|\, \mathbf{x}^i, \mathbf{\Theta}] \\
&= \arg\max_{\mathbf{\Theta}} \prod_{i=1}^{n} \left(\mathcal{N}(\mu^i, \sigma^2) \prod_{j=1}^{L} \sigma(\mathbf{w}_j^T \mathbf{x}^i)^{S_j^i} * [1 - \sigma(\mathbf{w}_j^T \mathbf{x}^i)]^{1-S_j^i} \right)
\end{aligned}$$

Taking log and solving for MLE(using first order derrivative for μ and σ) for this objective function; we get:

$$\begin{aligned}
\mu_i &= \sum_{j|S_j^i=1} C_j \\
\sigma_i^2 &= \frac{\sum_{i=1}^{n} (x^i - \mu^i)^2}{n}
\end{aligned}$$

The $\mathbf{w}_j$s can be obtained by performing logistic regression in $O(nd)$ time.

subPart4:

The soft Alternating optimization algorithm is as follows:

1. For $i \in [n]$ create 2^L copies of datapoint x^i.
   - Let $x^i \rightarrow \{x^{i,1}, x^{i,2}, ..., x^{i,2^L}\}$
2. Initialize $\mathbf{\Theta}^0$.
3. Update weights $\gamma^{i,k,t}$ using $\mathbf{\Theta}^0$

$$\begin{aligned}\gamma^{i,k,t} &= P[S^i | b^i, \mathbf{x}^i, \mathbf{\Theta}^t] \\ &= \frac{P[S^i = k \,|\, \mathbf{x}^i, \mathbf{\Theta}] \; P[b^i \,|\, x^i, S^i = k, \mathbf{\Theta}]}{\sum_{j=1}^{j=2^L} P[S^i = j \,|\, \mathbf{x}^i, \mathbf{\Theta}] \; P[b^i \,|\, x^i, S^i = j, \mathbf{\Theta}]} \\ &= \frac{\boldsymbol{\pi}_k^t \; \mathcal{N}(\mu_i, \sigma^2)}{\sum_{j=1}^{j=2^L} \boldsymbol{\pi}_j^t \; \mathcal{N}(\mu_i, \sigma^2)}\end{aligned}$$

   where $\pi_k = P[S^i = k | \mathbf{x}^i, \mathbf{\Theta}]$ which is the product of linear logistic models as described previously.

4. Update $\mathbf{\Theta}^{t+1}$,

$$\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} \; P[\{\mathbf{x}^{i,k}\}, \{b^{i,k}\}, \{\gamma^{i,k,t}\}]$$

   where $P[\mathbf{x}, b, \gamma \; | \mathbf{\Theta}] = P[x, b | \mathbf{\Theta}]^\gamma$.

Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning, 2017-18-a

Assignment Number: 2
Student Name: Abhisek Panda
Roll Number: 150026
Date: October 10, 2017

QUESTION 3

Given objective function is:

$$\begin{aligned} \underset{\mathbf{w},\{\xi_i\}}{\arg\min}\ & \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 \\ \text{s.t. } & y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1 - \xi_i, \text{ for all } i \in [n] \\ & \xi_i \geq 0, \text{ for all } i \in [n] \end{aligned} \qquad (P1)$$

subPart1:

To show that $\xi_i \geq 0$ are vacuous. Let ξ_i be negative, if possible, in the optimal solution obtained with corresponding objective value H. Observe the constraint $y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1 - \xi_i$. Since ξ_i is negative; $1 - \xi_i > 1 - 0 \implies y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle > 1\ (-0)$. Hence, we can always replace a negative ξ_i(if obtained) with a 0. Clearly, this reduces the objective function value to H', since $\xi_i^2 \geq 0\ \forall i$. This leads to a contradiction since we had assumed that the original H is the optimal value.

subPart2:

$$\begin{aligned} & y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1 - \xi_i \\ \implies & 1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0 \end{aligned}$$

and the constraint for ξ_i

$$\begin{aligned} & \xi_i \geq 0 \\ \implies & -\xi_i \leq 0 \end{aligned}$$

So, introduce parameters α_i and β_i so that the langrangian becomes:

$$L(\mathbf{w}, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\}) = \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \xi_i^2 + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle) - \sum_{i=1}^{n} \beta_i \xi_i \qquad (L1)$$

subPart3:

The objective is to solve

$$\underset{\mathbf{w},\{\xi_i\}}{\arg\min}\ \underset{\alpha_i \geq 0, \beta_i \geq 0}{\arg\max}\ L(\mathbf{w}, \{\xi_i\}, \{\alpha_i\}, \{\beta_i\})$$

Taking derrivative w.r.t $\mathbf{w}$, we get:

$$\begin{aligned} & 0 = 2\mathbf{w} + 0 + \sum_{i=1}^{n} \alpha_i * (-y^i \mathbf{x}^i) \\ \implies & \mathbf{w} = \frac{\sum_{i=1}^{n} \alpha_i y^i \mathbf{x}^i}{2} \end{aligned}$$

Similarly, taking derrivative w.r.t ξ_i, we get:

$$0 = 2\xi_i - \alpha_i - \beta_i \implies \xi_i = \frac{\alpha_i + \beta_i}{2}$$

Substituting these values in equation (L1), we get:

$$\begin{aligned} f(\{\alpha_i\}, \{\beta_i\}) &= \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (\mathbf{x}^i)^T\mathbf{x}^j + \frac{1}{4}\sum_{i=1}^{n}(\alpha_i+\beta_i)^2 + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y^i \left\langle \frac{1}{2}\sum_{j=1}^{n}\alpha_j y^j \mathbf{x}^j, \mathbf{x}^i \right\rangle \\ &\quad - \frac{1}{2}\alpha_i(\alpha_i+\beta_i) - \frac{1}{2}\beta_i(\alpha_+\beta_i) \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (\mathbf{x}^i)^T\mathbf{x}^j - \sum_{i=1}^{n}\frac{(\alpha_i+\beta_i)^2}{4} \end{aligned}$$

which is the required answer.

subPart 4:

Dual Problem for (P1) is given by:

$$f(\{\alpha_i\}, \{\beta_i\}) = \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\frac{(\alpha_i+\beta_i)^2}{4} - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (\mathbf{x}^i)^T\mathbf{x}^j$$

subject to the constraints: $\alpha_i \geq 0, \beta_i \geq 0$.
Note that the dual objective is maximized when $(\alpha_i + \beta_i)^2$ is as small as possible. Since β only appears in this term, we can conveniently set it to $0(\beta_i \geq 0)$. This implies that the original constraint of $\xi_i \geq 0$ was vacuous since it does not contribute to finding the optima.
Hence, the new Dual Problem becomes:

$$f(\{\alpha_i\}) = \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\frac{\alpha_i^2}{4} - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (\mathbf{x}^i)^T\mathbf{x}^j$$

subject to the constraints: $\alpha_i \geq 0$

Dual problem for original SVM is given by:

$$f(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j (\mathbf{x}^i)^T\mathbf{x}^j$$

subject to the constraints: $0 \leq \alpha_i \leq C$

Clearly, there is an extra term in order of $\sum_{i=1}^{n}\alpha_i^2$ due to taking ξ_i^2 instead of ξ_i in (P1). Moreover, there is an additional upper bound on α_i in the original SVM dual problem, which is due to the extra requirement of $\xi_i \geq 0$ there. No such constraint is to be satisfied in dual for (P1) as proved in subPart(1).