

Solution sketches for Assignment 1

COSC 6342 – Spring 2009

Question 1: Compare SL and RL

	Supervised Learning	Reinforcement Learning
Given	<ul style="list-style-type: none"> - Relies on training examples (data); correct and incorrect actions are known for each example. 	<ul style="list-style-type: none"> - Learning how to act intelligently in an initially unknown, possibly changing world. - Feedback is only obtained for actions that are taken, and feedback might be delayed; feedback is much more indirect in RL. - In contrast to SL, it deals with adaption and allows for changes.
Goal	<ul style="list-style-type: none"> - The goal is to correctly predict the output values of new data points x (static approach). 	<ul style="list-style-type: none"> - Deal with dynamic environments.

Question 2: Derive the solution for w_0 and w_1 at the bottom of page 30 of the textbook.

The error rate: $E(w_1, w_0 | X) = \sum_{t=1}^N (r^t - (w_1 x^t + w_0))^2$ gets minimum value at w_0, w_1 which make the partial derivatives of this rate to zero.

$$\frac{\partial E}{\partial w_0} = \sum_{t=1}^N 2(r^t - (w_1 x_t + w_0)) \cdot (-1) = -2 \sum_{t=1}^N (r^t - (w_1 x_t + w_0))$$

$$\frac{\partial E}{\partial w_0} = 0 \Leftrightarrow \sum_{t=1}^N (r^t - (w_1 x_t + w_0)) = 0$$

$$\Leftrightarrow \sum_{t=1}^N r^t - \sum_{t=1}^N w_1 x^t - N w_0 \Leftrightarrow w_0 = \frac{\sum_{t=1}^N r^t}{N} - w_1 \frac{\sum_{t=1}^N x^t}{N} = \bar{r} - w_1 \bar{x}$$

$$\begin{aligned}
\frac{\partial E}{\partial w_1} = 0 &\Leftrightarrow \sum_{t=1}^N (2x^t(w_0 + w_1x^t) - 2r^tx^t) = 0 \\
&\Leftrightarrow \sum_{t=1}^N (2x^t(\bar{r} - w_1\bar{x} + w_1x^t) - 2r^tx^t) = 0 \\
&\Leftrightarrow w_1 \sum_{t=1}^N ((x^t)^2 - x^t\bar{x}) + \sum_{t=1}^N (x^t\bar{r} - r^tx^t) = 0 \\
&\Leftrightarrow w_1 = \frac{\sum_{t=1}^N (r^tx^t - x^t\bar{r})}{\sum_{t=1}^N ((x^t)^2 - x^t\bar{x})} = \frac{\sum_{t=1}^N (r^tx^t) - \sum_{t=1}^N \frac{x^tr^t}{N}}{\sum_{t=1}^N ((x^t)^2) - \sum_{t=1}^N \frac{x^tx^t}{N}} = \frac{\sum_{t=1}^N (r^tx^t) - N\bar{x}\bar{r}}{\sum_{t=1}^N (x^t)^2 - N\bar{x}^2} \\
&\Rightarrow w_1 = \frac{\sum_{t=1}^N (r^tx^t) - N\bar{x}\bar{r}}{\sum_{t=1}^N (x^t)^2 - N\bar{x}^2}
\end{aligned}$$

Question 3: Solve the problem on transparency 9 of topics 3

$$R(a1 | x) = \lambda_{11} P(C1 | x) + \lambda_{12} P(C2 | x)$$

$$R(a2 | x) = \lambda_{21} P(C1 | x) + \lambda_{22} P(C2 | x)$$

$$\lambda_{12} = 9; \lambda_{21} = 90$$

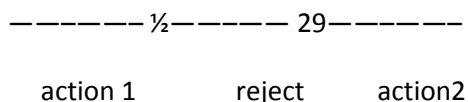
Denote $k = P(C2 | x) / P(C1 | x)$

a. Input: $P(C1 | x), P(C2 | x)$.

- Compute: $R(a1 | x) = 9 P(C2 | x)$;
- Compute: $R(a2 | x) = 90 P(C1 | x)$
- Decision making strategy:
 - o Choose action 1 if $R(a1 | x) < R(a2 | x) \Leftrightarrow 9 P(C2 | x) < 90 P(C1 | x) \Leftrightarrow k < 10$;
 - o Choose action 2: otherwise.

b. Input: $P(C1 | x), P(C2 | x), \lambda_{\text{reject}, 2} = 3; \lambda_{\text{reject}, 1} = 3$

- Compute: $R(a1|x) = 9P(C2|x)$
- Compute: $R(a2|x) = 90P(C1|x)$
- $R(\text{reject}|x) = \lambda_{\text{reject}, 2}P(C2|x) + \lambda_{\text{reject}, 1} \cdot P(C1|x) = 3 P(C2|x) + 3 P(C1|x)$
- Decision making strategy:
 - o Classify x to class 1 if $R(a1|x) < R(a2|x)$ and $R(a1|x) < R(\text{reject}|x) \Leftrightarrow k \leq 1/2$
 - o Classify x to class 2 if $R(a2|x) < R(a1|x)$ and $R(a2|x) < R(\text{reject}|x) \Leftrightarrow k \geq 29$
 - o Unable to classify x , choose action *reject* if $1/2 < k < 29$
- Or strategy:



Question 4: Formula 1.4 on page 10 of the Bishop uses regularization for determining the best weights in curve fitting. What is the motivation to use regularization? How is it used in the particular approach presented? What alternative methods could be used to accomplish the goals of regularization?

Motivation:

Motivation to use regularization: to avoid over-fitting, so can achieve the generality learning model.

How:

Adding a penalty term to the error function to avoid the coefficients from reaching large values. In the formula, the penalty term is sum of squares of all coefficients.

Alternative methods:

The other methods used in order to avoid over fitting problem are cross-validation and Bayesian curve fitting.

Question 6

***The old covariance matrix results in negative distance. Use the new covariance matrix:**

Compute *Mahalanobis* distance and *Euclidean* distance.

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 9 & -2 \\ 0 & -2 & 1 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{5} & \frac{2}{5} \\ 0 & \frac{2}{5} & \frac{9}{5} \end{bmatrix}$$

	\mathbf{x}	$\boldsymbol{\mu}$	$\mathbf{x} - \boldsymbol{\mu}$	$(\mathbf{x} - \boldsymbol{\mu})^T$	Mahalanobis Distance $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$	Euclidean Distance
Group 1	$[1, 1, 0]^T$	$[1, 0, 1]^T$	$[0, 1, -1]^T$	$[0, 1, -1]$	1.2	$\sqrt{2}$
	$[1, 1, 0]^T$	$[0, 1, 1]^T$	$[1, 0, -1]^T$	$[1, 0, -1]$	2.05	$\sqrt{2}$
	$[1, 0, 1]^T$	$[0, 1, 1]^T$	$[1, -1, 0]^T$	$[1, -1, 0]$	0.45	$\sqrt{2}$
Group 2	$[1, 0, 0]^T$	$[1, 1, -1]^T$	$[0, -1, 1]^T$	$[0, -1, 1]$	1.2	$\sqrt{2}$
	$[0, 1, 0]^T$	$[1, 1, -1]^T$	$[-1, 0, 1]^T$	$[-1, 0, 1]$	2.05	$\sqrt{2}$
	$[0, 0, -1]^T$	$[1, 1, -1]^T$	$[-1, -1, 0]^T$	$[-1, -1, 0]$	0.45	$\sqrt{2}$

How these Mahalanobis distance results differ from using Euclidean distance?

While all the Euclidean distance results are equal to each other and equal to $\sqrt{2}$, the Mahalanobis distance results of each group are different (0.45, 1.2 and 2.05).

Explain why particular pairs of vectors are closer/further away from each other when using Mahalanobis distance?

+ The Mahalanobis distance:

- The Mahalanobis distance from a test point to the center of mass is the Euclidean distance between them divided by the width of the ellipsoid in the direction of that test point. This width is defined as a standard deviation from the covariance matrix. $\sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}$
- Another form is square of the above Mahalanobis distance. This distance is used in our lecture slide (so I will follow this formula). $(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$
- In both two cases, Mahalanobis distance is proportional to Euclidean distance and inversely-proportional to the width of ellipsoid in the direction of the test point.

+ Let x_1, x_2, x_3 are three coordinates. x_2 has the highest variance, x_3 has the lowest variance; (x_1 and x_2), (x_1 and x_3) have covariance equaling to 0, (x_2 and x_3) has covariance differing from 0 but x_2 has too high variance (9) so that the widths of the ellipsoid in direction of x_2, x_3, x_1 are descending.

+ Look at the first group of 3 pairs:

- The Euclidean of them are equal.
- All different vectors of three pairs ($[0, 1, -1]$, $[1, 0, -1]$, $[1, -1, 0]$) has exactly one coordinate equals to 0 so they depend on other two coordinates.
- For the second pair, the different vector between two points has x_2 coordinate (direction with largest width) equal to 0, so it has the largest distance.
- For the third pair, the different vector between two points has x_3 coordinate (direction with smallest width) equal to 0, so it has the smallest distance.

+ Explanation for the second group is similar.

Advantages in using Mahalanobis distance of Euclidean distance

- Mahalanobis distance normalizes attributes based on variance; this makes all independent attributes equally important; and
- It alleviates problem caused by using different scales; and
- It downplays the impact on distance of corrected attributes.

Question 7: Give an example for which the model parameter estimation using maximum likelihood (ML) differs from the estimate obtained by using maximum a posterior (MAP).

Question 8 a) Derive the storage and run-time complexity of K-means based on the following input parameters:

k is the number of clusters

d is the number of attributes in the dataset

n is the number of objects to be clusters

t is the number of iterations K-means takes

Justify your result!

8b- Derive M-step

Group Daffodil's solution

To derive the M-step in equation 7.15.

Equation 7.12 says

$$\nabla_{\phi} \sum_t \sum_i h_t^i \log p(x^t | \Phi) = 0$$

For Gaussian components $\hat{p}_i(x^t | \Phi) \sim \mathcal{N}(m_i, S_i)$

$$p(x^t | \Phi) = \frac{1}{\sqrt{2\Pi} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}$$

Substituting M-Step is

$$\nabla_{\phi} \sum_t \sum_i h_t^i \log \left(\frac{1}{\sqrt{2\Pi} |\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)} \right) = 0$$

$$\sum_t \sum_i h_t^i \log \left(\frac{1}{\sqrt{2\Pi} |\Sigma_j|^{1/2}} \right) - \frac{1}{2} (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) = 0$$

$$\sum_t \sum_i h_t^i \left[-\frac{1}{2} (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) - \frac{1}{2} \log(2\Pi) - \frac{1}{2} \log|\Sigma_j| \right] = 0$$

Taking $-\frac{1}{2}$ common,

$\log(2\Pi)$ is a constant.

$$S = \Sigma_j$$

$$= \min_{m_i, S} \sum_t \sum_i h_t^i \left[(x-\mu_i)^T S^{-1} (x-\mu_i) - \log S \right]$$

When samples are small S may be singular and the inverses may not exist or |S| may be non zero but very small.

Therefore for the M-Step equation in case of shared covariance matrix becomes

$$= \min_{m_i, s} \sum_t \sum_i h_t^i \left[(x - \mu_i)^T S^{-1} (x - \mu_i) \right]$$

M Step in case of Shared diagonal matrix

$$\hat{p}_i(\mathbf{x}^t | \Phi) \sim \mathcal{N}(\mathbf{m}_i, s^2 \mathbf{I})$$

S is a shared diagonal matrix

$$\text{Starting with } \min_{m_i, s} \sum_t \sum_i h_t^i \left[(x - \mu_i)^T S^{-1} (x - \mu_i) \right]$$

$$(x - \mu_i)^T = (x - \mu_i)$$

S in this case is $I * s^2$

$$= \min_{m_i, s} \sum_t \sum_i h_t^i \left[\|x - \mu_i\|^2 (s^2)^{-1} \right]$$

$$= \min_{m_i, s} \sum_t \sum_i h_t^i \left[\frac{\|x - \mu_i\|^2}{s^2} \right]$$