Pattern Recognition and Machine Learning: Project 1, Zhengzuo Liu

Problem 1

Answer:

Proof: Lemma: In the view of the result of Marshall and Olkin (1960) and using $S = \{w^T x \geq b\}$. we obtain:

$\sup_{D_n} P_{x \sim D_n}(w^T x \geq b) = \frac{1}{1+d^2}$. with $d^2 = \inf_{x: w^T x \geq b} (x-\mu_n)^T \Sigma_n^{-1} (x-\mu_n)$.

①. If $w^T \mu_n \geq b$, take $x = \mu_n \Rightarrow d^2 = 0$. certainly the minimum.

②. If $w^T \mu_n \leq b$. write this as $d^2 = \inf_{d^T u \geq g} u^T u$. where $u = \Sigma_n^{-1/2}(x-\mu_n)$.

$d^T = w^T \Sigma_n^{1/2}$ and $g = b - w^T \mu_n \geq 0$. form the Lagrangian:

$$L(u, \lambda) = u^T u + \lambda(g - d^T u).$$

solve the optimum derivative, get

$$d^2 = \inf_{w^T x \geq b} (x-\mu_n)^T \Sigma_n^{-1} (x-\mu_n) = \frac{(b - w^T \mu_n)^2}{w^T \Sigma_n w}.$$

The the result is easily deducted.

Problem 2

Answer:

Prove: $K(\varepsilon) = \sqrt{\frac{1-\varepsilon}{\varepsilon}} = \sqrt{\frac{1}{\varepsilon} - 1}$. then $\min \varepsilon \Leftrightarrow \max K$.

$$-b + w^T \mu_p \geq K(\varepsilon) \sqrt{w^T \Sigma_p w} \Leftrightarrow w^T \mu_n - K(\varepsilon) \sqrt{w^T \Sigma_p w} \geq b$$

$$b - w^T \mu_p \geq K(\varepsilon) \sqrt{w^T \Sigma_p w} \Leftrightarrow b \geq w^T \mu_p + K(\varepsilon) \sqrt{w^T \Sigma_p w}$$

Therefore, original problem $\Leftrightarrow$

$$\max_{K, w, b} K \quad \text{s.t.} \quad w^T \mu_p - K\sqrt{w^T \Sigma_p w} \geq b \geq w^T \mu_n + K\sqrt{w^T \Sigma_n w} \qquad (8)$$

Since K is irrelevant to b, original problem $\Leftrightarrow$

$$\max_{K, w} K \quad \text{s.t.} \quad w^T \mu_p - K\sqrt{w^T \Sigma_p w} \geq w^T \mu_n + K\sqrt{w^T \Sigma_n w} \qquad (9).$$

$$\Leftrightarrow \max_{K, w} K \quad \text{s.t.} \quad \frac{w^T(\mu_p - \mu_n)}{K} \geq \sqrt{w^T \Sigma_p w} + \sqrt{w^T \Sigma_n w}$$

This implies $w^T(\mu_p - \mu_n) \geq 0$. otherwise the problem does not have a meaningful solution.

Then, we can set $w^T(\mu_p - \mu_n) = 1$ without loss of generality.

we get the original problem $\Leftrightarrow$

$$\max_{K, w} K \quad \text{s.t.} \quad \frac{1}{K} \geq \sqrt{w^T \Sigma_p w} + \sqrt{w^T \Sigma_n w}, \qquad (10)$$

$$w^T(\mu_p - \mu_n) = 1. \qquad (11).$$

Naturally, equals to

$$\min_{w} \sqrt{w^T \Sigma_p w} + \sqrt{w^T \Sigma_n w} \quad \text{s.t.} \quad w^T(\mu_p - \mu_n) = 1.$$

Then $b^* = w^{*T} \mu_p - \dfrac{\sqrt{w^{*T} \Sigma_p w^*}}{\sqrt{w^{*T} \Sigma_p w^*} + \sqrt{w^{*T} \Sigma_n w^*}} = w^{*T} \mu_n + \dfrac{\sqrt{w^{*T} \Sigma_n w^*}}{\sqrt{w^{*T} \Sigma_p w^*} + \sqrt{w^{*T} \Sigma_n w^*}}$

Problem 3

Answer:

The CVXPY package is adopted for solve the SOCP optimization problem. The complete code for this problem is presented in "main.py". The results are as follows:

Table 1. Accuracy and Run Time of Different Methods on Different Datasets

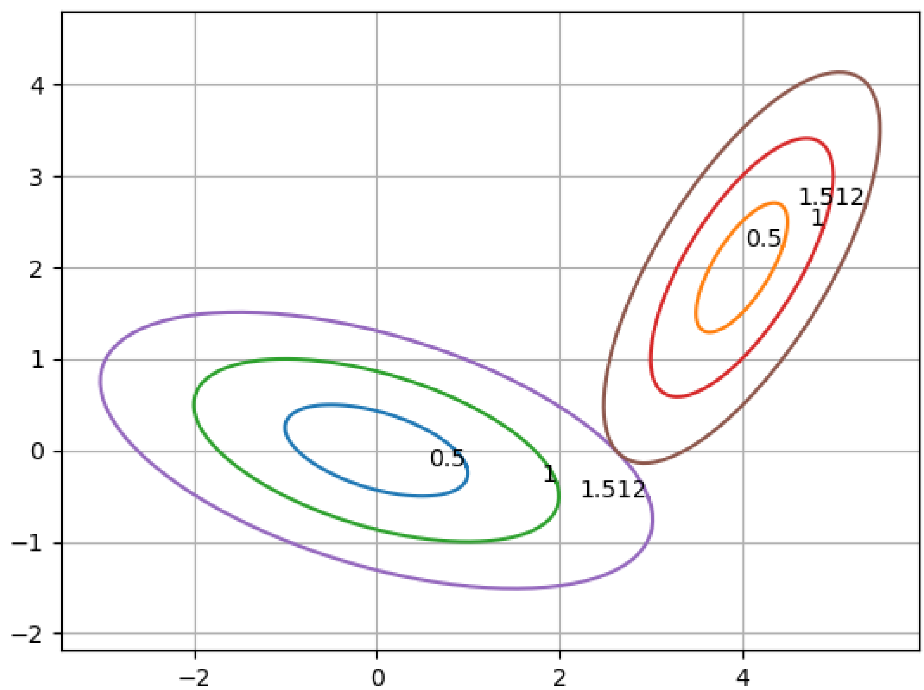
Dataset/Method	MPM	LDA	LR	SVM
Cancer	96.67% / 0.17s	95.94% / 0.04s	95.94% / 0.10s	96.96% / 0.06s
Diabetes	76.10% / 0.16s	77.79% / 0.04s	75.45% / 0.25s	75.97% / 0.15s
Sonar	72.38% / 0.26s	75.71% / 0.12s	82.86% / 0.10s	79.52% / 0.04s

It seems that MPM method is a little bit slower.

The average MPM's guaranteed error ϵ is 15.45%, 68.53%, and 35.24% for the three datasets. This result is consistent with that in paper [1], since $\epsilon = 1 - \alpha$ (α is defined in paper [1]).

Problem 4

Answer:



It is seen by the graph drawn above, that the smallest κ^* that make the two ellipsoids overlap is 1.512.

As calculated by MPM, the solution of (9) is also 1.512, being equivalent to κ^*. The complete code of the drawing and MPM calculation is presented in "Problem 4.py".

MPM is motivated theoretically as the minimization of a worst-case misclassification probability. However, MPM "happens to" have a very similar optimization problem form with FDA. Both optimization problems involves seeking for an optimized direction to separate two projected sets with small projected variances, with only minor difference in the way the covariance are combined. Then, by simply sum of squares inequality, the upper bound on the generalization error for FDA is found. In addition, when the covariance are equal, it is easily deducted that FDA and MPM provide identical classifiers. In a word, both similarity and difference between the two algorithms are profound and non-negligible.