Introduction to Machine Learning (67577)
Exercise III – Classification
Hadar Sharvit – 208287599

Theoretical Questions

Bayes Optimal and LDA

1. Show that $h_D = \underset{y\in\{\pm1\}}{\text{argmax}}(\Pr(\text{x}|\text{y})\Pr(y))$:
   We know from bayes theorem that $P(x|y)P(y) = P(y|x)$, therefore it is sufficient to show that $h_D = \underset{y\in\{\pm1\}}{\text{argmax}}\big(P(y|x)\big) = \underset{y}{\text{argmax}}(\{P(y=1|x), P(y=-1|x)\})$.
   Suppose in the first case that $\Pr(y=1|x) \geq 1/2$. If so, $\Pr(y=-1|x) < 1/2$ thus the corresponding argmax would be $y=+1$ (as $\Pr(y=1|x) > \Pr(y=-1|x)$). For the other case, $\Pr(y=1|x) < 1/2$, which in turn means $\Pr(y=-1|x) \geq 1/2$, so argmax would be $y=-1$. We can write those in compart form as followed
   $$\underset{y\in\{\pm1\}}{\text{argmax}}(\Pr(\text{x}|\text{y})\Pr(y)) = \underset{y\in\{\pm1\}}{\text{argmax}}(\text{P}(\text{y}|\text{x})) = \begin{cases} +1, & \Pr(y=1|x) \geq 1/2 \\ -1, & otherwise \end{cases} = h_D(x)$$

2. Show that $h_D(x) = \underset{y\in\{\pm1\}}{\text{argmax}}\ \delta_y(x)$ where $\delta_y(x) = x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln\Pr(y)$:
   From Q.1 we have $h_D(x) = \underset{y\in\{\pm1\}}{\text{argmax}}(\Pr(\text{x}|\text{y})\Pr(y))$, and we know that for calculating argmax, taking $\log h_D$ will provide the same result. If so, let us concentrate on the following:
   $$\ln\Pr(\text{x}|\text{y})\Pr(y) = \ln\Pr(x|y) + \ln\Pr(y)$$
   Substituting $f(x|y)$ we have
   $$\begin{aligned}\ln\Pr(x|y) &= \ln\left(\frac{1}{\sqrt{(2\pi)^d\det\Sigma}}\exp\left(-\frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)\right)\right) \\ &= -\ln\left(\frac{1}{\sqrt{(2\pi)^d\,det\,\Sigma}}\right) - \frac{1}{2}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y)\end{aligned}$$
   Since the first $\ln$ is not a function of y, it will not affect the result of argmax. Let us simplify the second component
   $$\begin{aligned}(x-\mu_y)^T\Sigma^{-1}(x-\mu_y) &= (x-\mu_y)^T(\Sigma^{-1}x - \Sigma^{-1}\mu_y) = (x^T-\mu_y^T)(\Sigma^{-1}x-\Sigma^{-1}\mu_y) \\ &= x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_y - \mu_y^T\Sigma^{-1}x + \mu_y^T\Sigma^{-1}\mu_y\end{aligned}$$
   Since Σ is the same for both ±1, $x^T\Sigma^{-1}\mu_y = \mu_y^T\Sigma^{-1}x$:
   $$= x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_y + \mu_y^T\Sigma^{-1}\mu_y$$
   Yet again, $x^T\Sigma^{-1}x$ is not a function of y, therefore is not relevant for the calculation of argmax. In conclusion we have
   $$\underset{y\in\{\pm1\}}{\text{argmax}}(\ln P(x|y)) = x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y$$
   And consequently

$$h_D(x) = \underset{y\in\{\pm1\}}{\text{argmax}}(\Pr(\text{x}|\text{y})\Pr(y)) = \underset{y\in\{\pm1\}}{\text{argmax}}\left(x^T\Sigma^{-1}\mu_y - \frac{1}{2}\mu_y^T\Sigma^{-1}\mu_y + \ln\Pr(y)\right) = \underset{y\in\{\pm1\}}{\text{argmax}}\ \delta_y(x)$$

3. Write your formula for estimating $\mu_{\pm 1}$, Σ and $\Pr(y)$ based on $S = \{(x_i, y_i)\}_{i=1}^m$:
   We can calculate the mean by summing the ratio of occurrences:
   $$\mu_{\pm 1} = \text{mean}(x[y = \pm 1]) = \frac{\sum x_i[y = \pm 1]}{\sum 1[y = \pm 1]}$$
   The probability would be the sum of occurrences divided by the total size
   $$\Pr(y) = \frac{1}{m} \sum 1[y_i = y]$$
   Similar to what we have seen in lecture 1, we can write
   $$\Sigma_{\pm} = \text{cov(x)} = \frac{1}{m-1} \sum_{y=\pm 1} (x[y = \pm 1] - \widehat{\mu_{\pm}})(x[y = \pm 1] - \widehat{\mu_{\pm}})^T$$
   And $\Sigma = \Sigma_+ + \Sigma_-$

Type I errors

4. Let us write the possible cases, given that $y = 1$ indicates spam and $y = -1$ is non-spam:
   1. If the current mail is non-spam ($y = -1$):
      a. $\hat{y} = -1$ (true negative): I have correctly identified the mail as non spam
      b. $\hat{y} = 1$ (false positive): I have declared the mail is spam (which is wrong)
   2. If the current mail is spam ($y = 1$):
      a. $\hat{y} = 1$ (true positive): I have correctly identified the mail as spam
      b. $\hat{y} = -1$(false negative): I have declared the mail is non-spam (which is wrong)

   1.b is denoted a Type-I error – we would wish to avoid declaring regular mail as spam

   2.b is not as bad, though still an error.

SVM – Formulation

5. Write the Hard-SVM problem as a QP problem:
   We can set $Q = 2\mathbb{I}_n, a = \vec{0}_n$, so the QP takes the form of
   $$\underset{\text{v}\in\mathbb{R}^{\text{n}}}{\text{argmin}}\ v^T v$$
   $$s.t\ Av \leq d$$
   Since $\big|\big|w\big|\big| \geq 0$, $\text{argmax}\ \big|\big|\text{w}\big|\big|^2 = \text{argmax}\ \big|\big|\text{w}\big|\big|$, so for $v := w$ we have the QP argmin set. (as $v^T v = ||w||$)
   Simplifying the conditions:
   $$y_i \langle w, x_i \rangle + y_i b \geq 1 \rightarrow \langle w, x_i \rangle \geq \frac{1}{y_i} - b \overset{(*)}{=} y_i - b$$
   $$-\langle w, x_i \rangle \leq b - y_i$$
   Or in matrix form
   $$= -(w_1 \quad \dots \quad w_n) \begin{pmatrix} | \\ x_i \\ | \end{pmatrix} \leq \begin{pmatrix} b - y_1 \\ \vdots \\ b - y_n \end{pmatrix}$$
   $$= -(- \quad x_i \quad -) \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \leq \begin{pmatrix} b - y_1 \\ \vdots \\ b - y_n \end{pmatrix}$$

   Therefore, in matrix form for every i:

$$-\underbrace{\begin{pmatrix} - & x_1 & - \\ \vdots & \vdots & \vdots \\ - & x_m & - \end{pmatrix}}_{A}\underbrace{\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}}_{v} \leq \underbrace{\begin{pmatrix} b - y_1 \\ \vdots \\ b - y_n \end{pmatrix}}_{d}$$

This finishes the transition to QP

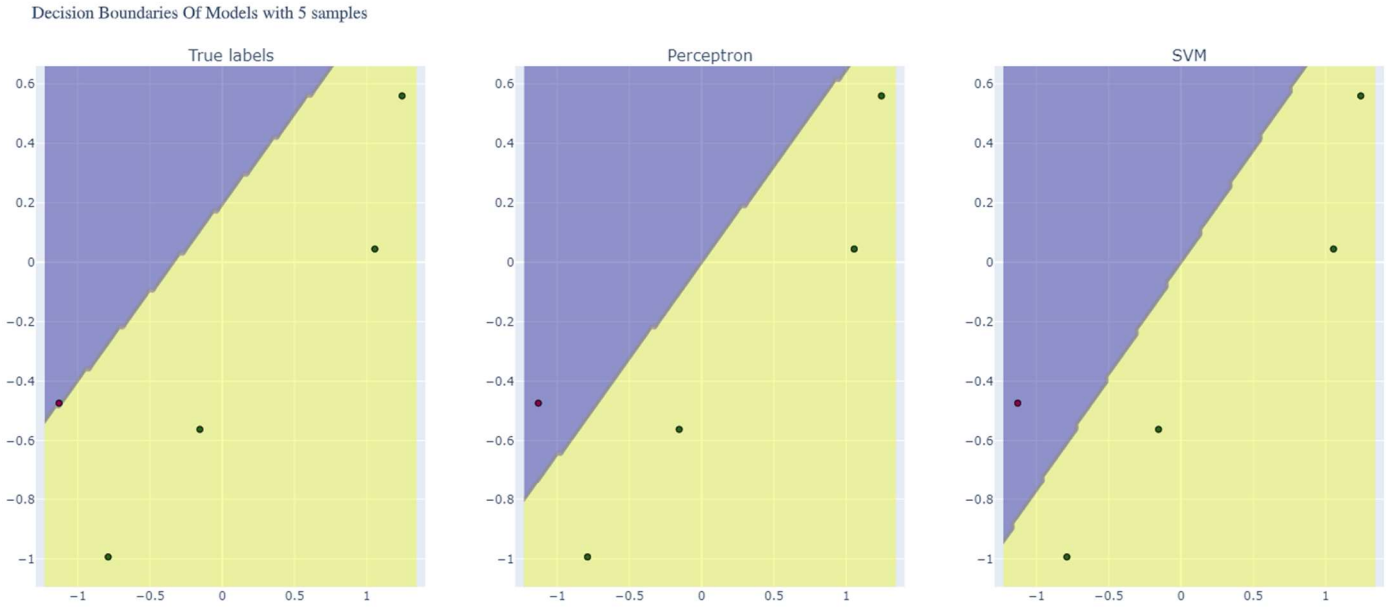
(*) $y_i \in \{\pm 1\} \rightarrow 1/y_i = y_i$

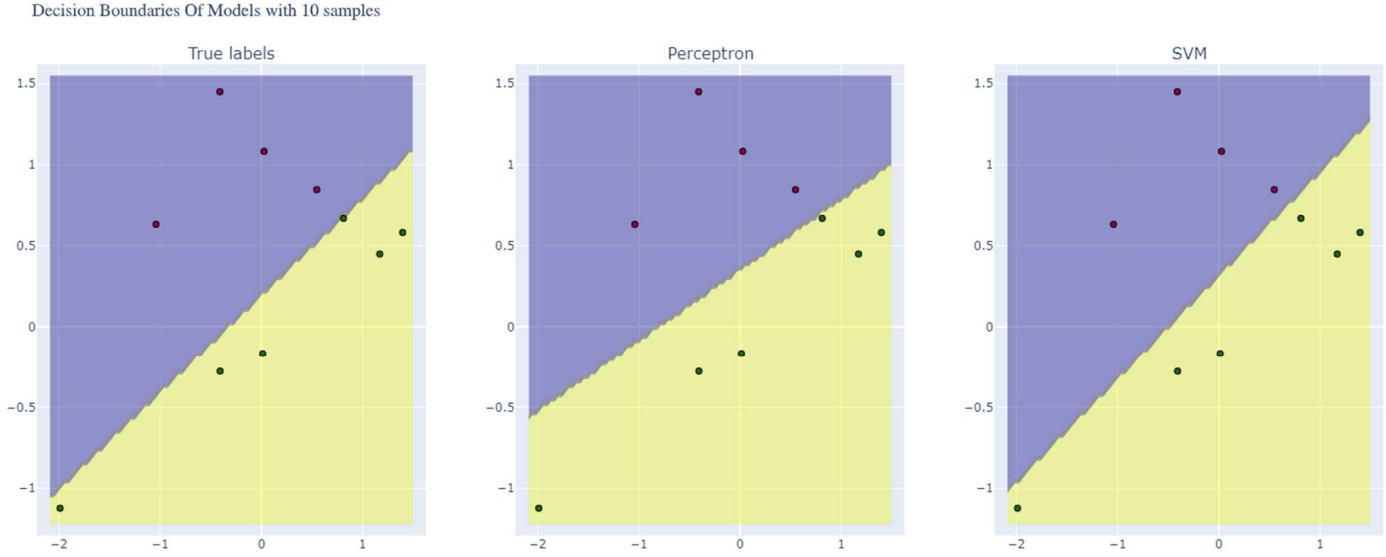
6.

Practical Questions

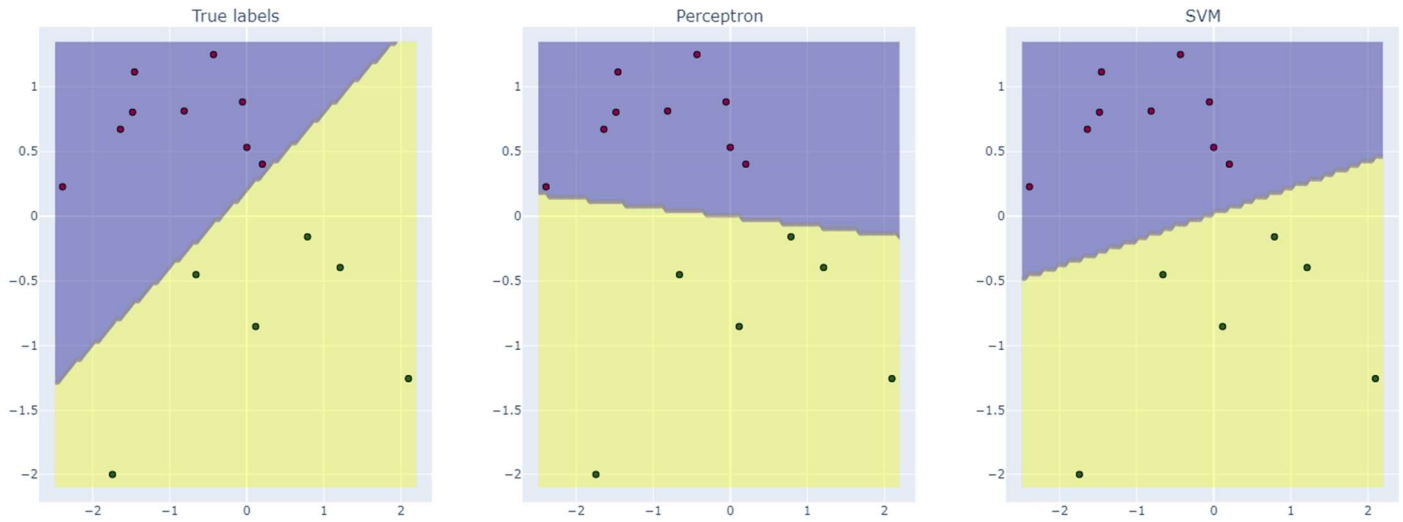
Implemention and simulation-comparison of different classifiers

7. Code
8. Code
9. Here are the plots decision boundaries of models with m samples:

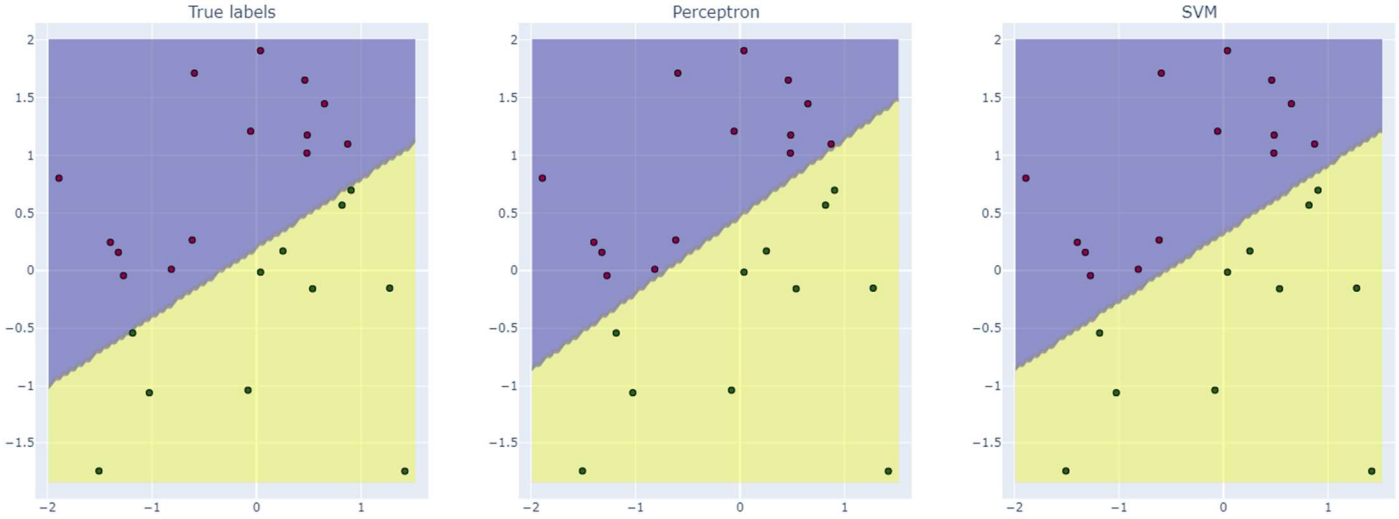




Decision Boundaries Of Models with 15 samples



Decision Boundaries Of Models with 25 samples



Decision Boundaries Of Models with 70 samples