Name: Aditya Agrawal

Entry No.: 2021AM10198

Date________ Page No.________

APL405

HOMEWORK-3

Q1. Given,

$$\min_{\underline{\theta}, b, \xi_i} \|\underline{\theta}\|_2^2 + C\sum_{i=1}^{n} \xi_i^2 \quad \text{s.t. } y_i(\underline{\theta}^T \underline{x}_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \ \forall i$$

We define the following Lagrangian:

$$\mathcal{L}(\underline{\theta}, b, \underline{\xi}, \underline{\alpha}, \underline{\gamma}) = \|\underline{\theta}\|^2 + C\sum_{i=1}^{N} \xi_i^2 + \sum_{i=1}^{N} \alpha_i \left(1 - \xi_i - y_i(\underline{\theta}^T \underline{x}_i + b)\right) - \sum_{i=1}^{N} \gamma_i \xi_i$$

$$\frac{\partial \mathcal{L}}{\partial \underline{\theta}} = 0 \Rightarrow 2\underline{\theta} - \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i = 0 \Rightarrow \underline{\theta} = \frac{1}{2}\sum_{i=1}^{N} \alpha_i y_i \underline{x}_i$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow -\sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow 2C\xi_i - \alpha_i - \gamma_i = 0 \Rightarrow \xi_i = \frac{1}{2C}(\alpha_i + \gamma_i)$$

Now,

$$\mathcal{L} = \underline{\theta}^T \underline{\theta} + \frac{1}{4C}\sum_{i=1}^{N} (\alpha_i + \gamma_i)^2 + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \underline{\theta}^T \sum_{i=1}^{N} \alpha_i y_i \underline{x}_i - b\sum_{i=1}^{N} \alpha_i y_i - \sum \gamma_i \xi_i$$

$$= -\frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j + \frac{1}{4C}\sum_{i=1}^{N} (\alpha_i + \gamma_i)^2 + \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \frac{1}{2C}(\alpha_i + \gamma_i)^2$$

$$\therefore G(\underline{\alpha}, \underline{\gamma}) = -\frac{1}{4}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \underline{x}_i^T \underline{x}_j + \sum_{i=1}^{N} \alpha_i - \frac{1}{4C}\sum_{i=1}^{N} (\alpha_i + \gamma_i)^2$$

Thus, the dual SVM is the maximization (i.e. −minimization) of $G(\underline{\alpha}, \underline{\gamma})$.

$$\therefore \min_{\underline{\alpha}, \underline{\gamma}} \frac{1}{4}\sum\sum \alpha_i \alpha_j y_i y_j \underline{x}_i^T \underline{x}_j - \sum_{i=1}^{N} \alpha_i + \frac{1}{4C}\sum_{i=1}^{N} (\alpha_i + \gamma_i)^2$$

$$\text{s.t. } \sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \ \forall \ i = 1, \ldots, N$$

Q.3 We define model parameters as:

$$\mathbb{E}[Z] = \mu \quad \text{(Bias)}$$
$$\mathbb{E}[(Z-\mu)^2] = \sigma^2 \quad \text{(Variance)}$$
$$\cancel{\mathbb{E}}\ Cor(Z_i, Z_j) = \rho_{ij} \quad \text{(Correlation)}$$

Now,

For individual models,

$$\text{Bias} = \mathbb{E}[Z_b] = \mu \quad \{b = 1, 2, \dots, B\} \quad \{\because \text{Uniform sampling}\}$$
$$\text{Variance} = \mathbb{E}[(Z_b-\mu)^2] = \sigma^2 \quad \{b = 1, 2, \dots, B\}$$
$$\text{Correlation} = Corr(Z_i, Z_j)_{i \neq j} = \rho \quad \{i, j = 1, 2, \dots, B\}$$

Considering bagged models:

$$\text{Bias} = \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} Z_b\right] = \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}[Z_b] = \frac{\cancel{B}\times\mu}{\cancel{B}} = \mu$$

$\Rightarrow$ Bias remains the same in both models.

$$\text{Variance} = \mathbb{E}\left[\left(\frac{1}{B}\sum_{b=1}^{B} Z_b\right)^2\right] - \left(\mathbb{E}\left[\sum_{b=1}^{B}\frac{1}{B}Z_b\right]\right)^2$$

$$\mathbb{E}\left[\left(\frac{1}{B}\sum_{b=1}^{B} Z_b\right)^2\right] = \frac{1}{B^2}\mathbb{E}\left[\sum Z_b^2 + \cancel{2}\sum_{i\neq j,\ i,j=1,\dots,B} Z_i Z_j\right]$$
$$= \frac{1}{B^2}\sum_b \mathbb{E}[Z_b^2] + \frac{\cancel{2}1}{B^2}\mathbb{E}\left[\sum Z_i Z_j\right]$$
$$= \frac{B(\sigma^2+\mu^2)}{B^2} + \frac{(B-1)B}{B^2}\left(Cov(Z_i, Z_j) + \mu^2\right)$$
$$= \frac{\sigma^2+\mu^2}{B} + \frac{(B-1)(\rho\sigma^2+\mu^2)}{B}$$
$$= \frac{\sigma^2(\rho(B-1)+1) + B\mu^2}{B}$$

$$\therefore \text{Variance} = \frac{\sigma^2(\rho(B-1)+1)}{B} \cancel{+\mu^2} - \cancel{\mu^2} = \frac{\sigma^2}{B} + \rho\sigma^2\left(1 - \frac{1}{B}\right)$$

Date
Page No.

Since $\rho \leq 1$

$$\Rightarrow \rho\sigma^2\left(1-\frac{1}{B}\right) \leq \sigma^2\left(1-\frac{1}{B}\right)$$

$$\Rightarrow \rho\sigma^2\left(1-\frac{1}{B}\right) + \frac{\sigma^2}{B} \leq \sigma^2$$

Clearly, the variance for bagging model is lesser than the individual model.
The bias - variance decomposition suggests:

$$\mathbb{E}[(y-\hat{y})^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible error}$$

Thus, the bias is unchanged while variance reduces in bagging resulting in a net reduction of the squared error loss.

Q.4. Given,

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)} \left(\ln \mathcal{N}(x^{(i)} \mid \underline{\mu}_m, \underline{\underline{\Sigma}}_m) + \ln \pi_m\right)$$

Let $J(\theta) = \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)} \left(\ln \{\mathcal{N}(\underline{x}^{(i)} \mid \underline{\mu}_m, \underline{\underline{\Sigma}}_m)\} + \ln \pi_m\right)$ $\{\underline{\theta} = \{\underline{\mu}_m, \underline{\underline{\Sigma}}_m, \pi_m\}\}$

For maxima, $\frac{\partial J}{\partial \theta} = 0$

We know the estimated $\hat{\underline{\mu}}_m$ for the current M-steps, thus we only need to update $\underline{\underline{\Sigma}}_m$.

$$\therefore \frac{\partial J}{\partial \underline{\underline{\Sigma}}_m} = 0$$

$$J = \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)} \left(\ln\left(\frac{1}{(2\pi)^{p/2}|\underline{\underline{\Sigma}}_m|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_m)^T \underline{\underline{\Sigma}}_m^{-1}(\underline{x}-\underline{\mu}_m)}\right) + \ln \pi_m\right)$$

$$= \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)} \left[-\frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(|\underline{\underline{\Sigma}}_m|) - \frac{1}{2}(\underline{x}_i-\underline{\mu}_m)^T \underline{\underline{\Sigma}}_m^{-1}(\underline{x}_i-\underline{\mu}_m) + \ln \pi_m\right]$$

Let $\Lambda_m = \Sigma_m^{-1} \Rightarrow |\Lambda_m| = \dfrac{1}{|\Sigma_m|}$

$$\therefore J = \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)}\left(-\frac{p}{2}\ln(2\pi) + \frac{1}{2}\ln(|\Lambda_m|) - \frac{1}{2}(x-\mu_m)^T\Lambda_m(x-\mu_m) + \ln\pi_m\right)$$

$$\frac{\partial J}{\partial \Sigma_m} = \frac{\partial J}{\partial \Lambda_m}\frac{\partial \Lambda_m}{\partial \Sigma_m}$$

$\Rightarrow$ ~~Maximi~~ $\dfrac{\partial J}{\partial \Lambda_m} = 0 \Rightarrow \dfrac{\partial J}{\partial \Sigma_m} = 0$ since $\dfrac{\partial \Lambda_m}{\partial \Sigma_m}$ is non-zero $\{\Sigma_m^{-1}$ exists

$$\therefore \frac{\partial J}{\partial \Lambda_m} = \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)}\left(\frac{1}{2}\frac{1}{|\Lambda_m|}\,\mathrm{adj}(\Lambda_m) - \frac{1}{2}(x-\mu_m)(x-\mu_m)^T\right) = 0$$

{Using identities: $\dfrac{d|A|}{dA} = \dfrac{\mathrm{adj}(A)}{|A|}$ & $\dfrac{d(x^TAx)}{dA} = xx^T$ }

Also, we know that Σ_m and thus Λ_m are all equal.

$$\Rightarrow 0 = \frac{1}{2}\frac{\mathrm{adj}(\Lambda_m)}{|\Lambda_m|}\sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)} - \frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)}(x-\mu_m)(x-\mu_m)^T$$

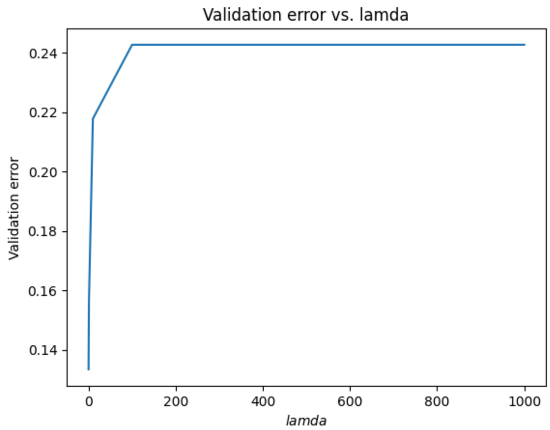
$$\Rightarrow \frac{\mathrm{adj}(\Lambda_m)}{|\Lambda_m|}\sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)} = \sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)}(x-\mu_m)(x-\mu_m)^T$$

$$\Lambda_m^{-1} = \frac{\mathrm{adj}(\Lambda_m)}{|\Lambda_m|} = \Sigma_m$$

$$\therefore \Sigma_m = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)}(x-\mu_m)(x-\mu_m)^T}{\sum_{i=1}^{N}\sum_{m=1}^{M} w_m^{(i)}}$$

Q2. (b) Classification accuracy on training set: 79.515%
Classification accuracy on testing set: 79.66%

(c)



(d) Accuracy on training dataset: 87.865%; Accuracy on testing dataset: 87.57%