Q.2 For 2 class case the formulation used in lecture notes is as follows:

$$\hat{\omega} = \underset{\bar{\omega}}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i \ln(h(x_i;\omega)) + (1-y_i)\ln(1-h(x_i;\omega))$$

The loss function is interpreted as if y_i is class I, then the loss is $y_i \ln(h(x_i;\omega))$ and otherwise y_i is the reference class (class 0) whose loss is $(1-y_i)\ln(1-h(x_i;\omega))$.

This is extended to C class classification as follows. Consider a C-1 dim 1 hot vector $y_{i,k}$ s.t: $y_i = \mathbb{1}(y_i = k)$, $k = 1, 2, \cdots, C-1$.

$$y_{i,k} = \begin{cases} 1, & \text{if } y_i = k \\ 0, & \text{if } y_i \neq 0 \end{cases}, \quad k = 1, 2, \cdots, C-1$$

No indicator is introduced for class 0, $y_i = \bar{0}$ for class 0.

Then the multiclass formulation is:

$$\hat{\omega}_c = \underset{\bar{\omega}_c}{\operatorname{argmin}} - \sum_{i=1}^{N} \sum_{k=1}^{C-1} y_{ik} \log(P(y_i = k|\bar{x}_i)) + (1 - \sum_{k=1}^{C-1} y_{ik}) \log P(y_i = 0|\bar{x}_i) + \frac{\lambda}{2} \sum_{k=1}^{C-1} \|\bar{\omega}_k\|_2^2$$

For learning $\bar{\omega}_c = \bar{\omega}_0 - \rho \frac{\partial E(\bar{\omega}_k)}{\partial \bar{\omega}_c}$

$$\frac{\partial E(\bar{\omega}_c)}{\partial \bar{\omega}_c} = \frac{\partial}{\partial \bar{\omega}_c} \left\{ \sum_{i=1}^{N} \sum_{k=1}^{C-1} y_{ik} \log(P(y_i = k|\bar{x}_i)) + (1 - \sum_{k=1}^{C-1} y_{ik}) \log P(y_i = 0|\bar{x}_i) \right\} + \frac{\partial}{\partial \bar{\omega}_c} \left(\frac{\lambda}{2} \sum_{k=1}^{C-1} \|\bar{\omega}_k\|_2^2 \right)$$

$$= \underbrace{-\sum_{i=1}^{N} \frac{\partial}{\partial \bar{\omega}_c} \left\{ \sum_{k=1}^{C-1} y_{ik} \log(P(y_i = k|\bar{x}_i)) \right\}}_{\text{Part A}} + \underbrace{\frac{\partial}{\partial \bar{\omega}_c} \left\{ (1 - \sum_{k=1}^{C-1} y_{ik}) \log P(y_i = 0|\bar{x}_i) \right\}}_{\text{Part B}} + \lambda \bar{\omega}_c^T$$

$$\text{Part A} = \frac{\partial}{\partial \bar{\omega}_c} \sum_{k=1}^{C-1} y_{ik} \log(P(y_i = k|\bar{x}_i)) = \sum_{k=1}^{C-1} y_{ik} \underbrace{\frac{\partial}{\partial \bar{\omega}_c} \log(P(y_i = k|\bar{x}_i))}_{\text{Part C}}$$

Using $P(y_i = k|\bar{x}_i) = \dfrac{\exp(-\bar{\omega}_k^T \bar{x}_i)}{1 + \sum_{k=1}^{C-1} (-\bar{\omega}_k^T \bar{x}_i)}$

$$\text{Part C} = \frac{1}{P(y_i = k|\bar{x}_i)} \cdot \frac{\partial}{\partial \bar{\omega}_c} \left(\frac{\exp(-\bar{\omega}_k^T \bar{x}_i)}{1 + \sum_{j=1}^{C-1} \exp(-\bar{\omega}_j^T \bar{x}_i)} \right)$$

Page No.
Date:

Part C = $\frac{1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_i)}{\exp(-\omega_K^T \bar{x}_i)} \left[\frac{-\exp(-\omega_K^T \bar{x}_i)}{(1+\sum_{j=1}^{C-1}\exp(-\omega_j^T x_i))^2} \cdot \exp(-\omega_C^T x_i) \cdot -x_i \right.$

$\left. + \frac{\exp(-\omega_K^T \bar{x}_i) \cdot -x_i \cdot 1_{(K=C)}}{(1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_i))} \right]$

Note

$\frac{\partial \exp(-\omega_j^T x_i)}{\partial \omega_C} = \begin{cases} 0, & K \neq C \\ -\exp(-\omega_C^T \bar{x}_i)\bar{x}_i & \text{if } K=C \end{cases}$

Part C = $\left[\frac{\exp(-\omega_C^T \bar{x}_i)\bar{x}_i}{(1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_i))} + x_i 1_{(K=C)} \right]$

Part B = $\frac{\partial}{\partial \omega_C} \left[(1-\sum_{K=1}^{C-1} y_{iK}) \log(1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_i)) \right]$

$= -\left[1-\sum_{K=1}^{C-1} y_{iK}\right] \frac{\partial}{\partial \omega} \left[\log(1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_i)) \right]$

$= -(1-\sum_{K=1}^{C-1} y_{iK}) \frac{1}{1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_j)} \cdot \exp(-\omega_C^T \bar{x}_i) \cdot -\bar{x}_i$

$= \frac{(1-\sum_{K=1}^{C-1} y_{iK}) x_i \exp(-\omega_C^T x_i)}{1+\sum_{j=1}^{C-1}\exp(-\omega_j^T \bar{x}_i)} = x_i (1-\sum_{K=1}^{C-1} y_{iK}) P(y_i=C|x_i)$

Part A = $\sum_{K=1}^{C-1} y_{iK} \left[\frac{\exp(-\omega_C^T \bar{x}_i)\bar{x}_i}{1+\sum_{j=1}^{C-1}(-\omega_j^T \bar{x}_i)} - x_i 1_{(K=C)} \right]$

$= \sum_{iK=1}^{C-} \bar{x}_i y_{iK} \left[P(y_i=C|\bar{x}_i) - 1_{(K=C)} \right]$

Part A + Part C $= \bar{x}_i \sum_{K=1}^{C-1} y_{iK} \left[P(y_i=C|x_i) - 1_{(K=C)} \right] + \bar{x}_i (1-\sum_{K=1}^{C-1} y_{iK}) P(y_i=C|\bar{x}_i)$

$= \bar{x}_i \left[\sum_{K=1}^{C-1} y_{iK} P(y_i=C|x_i) - \sum_{K=1}^{C-1} 1_{(K=C)} y_{iK} + P(y_i=C|x_i) - \sum_{K=1}^{C-1} y_{iK} P(y_i=C|x_i) \right]$

$= \bar{x}_i \left(P(y_i=C|x_i) - \sum_{K=1}^{C-1} y_{iK} 1_{(K=C)} \right] = x_i \left[P(y_i=C|x_i) - y_{iC} \right]$

Substituting back in learning equation:

$$\frac{\partial E(w)}{\partial \bar{w}_c} = -\sum_{i=1}^{N} x_i \left[P(y_i = c \mid x_i) - y_{ic} \right] + \lambda \bar{w}_c$$

Thus the learning rule with learning rate ρ is:

$$\bar{w}_{c,t+1} \leftarrow \bar{w}_{c,t} + \rho \left(\sum_{i=1}^{N} x_i \left[P(y=c \mid x_i) - y_{ic} \right] - \lambda \bar{w}_c \right)$$

This rule is different from lecture notes which is

$$\bar{w}_{t+1} \leftarrow w_t + \rho \left(\sum_{i=1}^{N} x_i (y_i - h(x_i; w)) - \lambda \bar{w}_t \right)$$

This is because in lecture the probability of class 0 is defined as $P(y=0 \mid \bar{x}) = \frac{\exp(-w^T x)}{1 + \exp(-w^T x)}$

whereas in this problem, the probability of class 0 is:

$$P(y_i = 0 \mid x_i) = \frac{1}{1 + \sum_{k=1}^{C-1} \exp(-w_k^T x_i)}$$

This subtle difference introduces a negative sign.

Q2:

Part 2)

Reporting Training and validation for 3 Fold Cross-validation for linear kernel with varying C. For other parameters, default values were used, which were as follows:

```
        degree=3,
        gamma="scale",
        coef0=0.0,
        shrinking=True,
        probability=False,
        tol=1e-3,
        cache_size=200,
        class_weight=None,
        verbose=False,
        max_iter=-1,
        decision_function_shape="ovr",
        break_ties=False,
        random_state=None,
```

C = 0.01	C = 0.05	C = 0.1	C = 0.5	C = 1.0
Train: 0.8454 Val: 0.8448	Train: 0.8481 Val: 0.8472	Train: 0.8491 Val: 0.8479	Train: 0.8498 Val: 0.8479	Train: 0.8499 Val: 0.8479

The Best linear model is with C = 1.0, which is to be expected, as a higher C value means that the classifier pays more weight on reducing error (increasing error) rather than maximising margin. I suspect that with an even higher C accuracy could be improved even more.

Part 3)

Reporting Training and Validtion Accuracy for rbf kernel with varying C and gamma. For other parameter, default values were used, same as Part 2.

γ \ C	**0.01**	**0.05**	**0.10**	**0.50**	**1.00**
0.01	Train: 0.7592 Val: 0.7592	Train: 0.8316 Val: 0.8310	Train: 0.8385 Val: 0.8379	Train: 0.8445 Val: 0.8436	Train: 0.8471 Val: 0.8451
0.05	Train: 0.8201 Val: 0.8201	Train: 0.8369 Val: 0.8356	Train: 0.8416 Val: 0.8397	Train: 0.8515 Val: 0.8448	Train: 0.8566 Val: 0.8470
0.10	Train: 0.8204 Val: 0.8195	Train: 0.8359 Val: 0.8342	Train: 0.8421 Val: 0.8390	Train: 0.8586 Val: 0.8464	Train: 0.8669 Val: 0.8472
0.50	Train: 0.7592 Val: 0.7592	Train: 0.7945 Val: 0.7889	Train: 0.8136 Val: 0.8055	Train: 0.8868 Val: 0.8328	Train: 0.9361 Val: 0.8347
1.00	Train: 0.7592	Train: 0.7592	Train: 0.7640	Train: 0.8380	Train: 0.9613

Val: 0.7592 Val: 0.7592 Val: 0.7619 Val: 0.7889 Val: 0.7975

Once again, the best validation accuracy was observed at C = 1.0. For gamma, too small is underfitting, too large is overfitting, though the best performance was observed at gamma = 0.01.

Part 4)

Based on the previous 2 parts, the best performance was for the linear kernel with C = 1.0. Although for C = 0.1, 0.5, 1.0, the validation accuracy did not change, the training accuracy was the highest for C = 1.0, so a linear kernel with C = 1.0 was chosen to train on the entire training set and used for test predictions.

	Linear Kernel With C = 1.0
Accuracy of SVM	Test: 0.8498, Train: 0.8499

Date:

Q.3 The optimization of soft margin SVM is as follows:

$$\min_{w,b,\xi_i} \frac{\|\bar{w}\|_2^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)$$

$$s.t.\ y_i(\bar{w}\cdot\bar{x}_i + b) \geq 1 - \xi_i \quad i = 1,\dots,N$$

$$\text{and } \xi_i \geq 0$$

The two constraints can be rewritten as follows:

$$\xi_i \geq 1 - y_i(\bar{w}_i \cdot \bar{x}_i + b) \text{ and } \xi_i \geq 0$$

This can be combined to form $\xi_i \geq \max(0, 1 - y_i(\bar{w}_i \cdot \bar{x}_i + b))$

Since the optimization formulation minimizes $\sum_{i=1}^{N} \xi_i$, the minimum must occur at equality, i.e

$$\xi_i = \max(0, 1 - y_i(\bar{w}_i \bar{x}_i + b)).$$

Which is the hinge loss function for SVMs. Substituting it back in objective gives:

$$\min_{w,b,\xi_i} \frac{\|\bar{w}\|_2^2}{2} + C \sum_{i=1}^{N} \max(0, 1 - y_i(\bar{w}_i \bar{x}_i + b))$$

So the optimization problem is hinge loss minimization + regularizer which is similar to empirical risk minimization. That is (from L5 Page 37):

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i;\theta), y_i) + \lambda \Omega(\theta)$$

If we consider this to be a linear problem, that is $\theta \in \mathbb{R}^n$ and $f(x_i;\theta) = \bar{w}\cdot\bar{x}_i + b$ and the loss to be hinge loss that is $\ell(f(x_i;\theta), y_i) = \max(0, 1 - y_i f(x_i;\theta)) =$

$$\max(0, 1 - y_i(\bar{w}\cdot\bar{x}_i + b)$$

and the penalty term to be L_2 regularizer that is $\Omega(\theta) = \Omega(w) = \frac{\|\bar{w}\|_2^2}{2}$, the risk minimization problem becomes the same as one for SVM

$$\hat{w} = \arg\min_{\bar{w}} \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - y_i(\bar{w}_i \bar{x} + b)) + \lambda \frac{\|w_2\|^2}{2}$$

where $C = \frac{1}{\lambda N}$ and the SVM formulation is multiplied by λ.

The presence of a scalar positive multiple does not change the optimization formulation, though the optimal value will be scaled.

AI 6102: Individual Assignment

For ANSHUL YADAV

Q4 The formulation for linear regression from L3 (Page 47) is:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (\bar{w} \cdot \bar{x}_i - y_i)^2 + \frac{\lambda}{2} \|\bar{w}\|_2^2$$

To extend this for non-linear cases, one needs to use feature map $\phi(x): \mathbb{R}^M \to \mathbb{R}^H$ where $H \gg M$. Therefore the new formulation becomes:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (\bar{w} \cdot \overline{\phi(x_i)} - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2$$

where $\bar{w}$ is a H-dim vector so that the dot product is correctly defined.

To find a closed form solution $\frac{\partial \hat{w}}{\partial \bar{w}} = 0$

$$\frac{\partial}{\partial \bar{w}} \left(\frac{1}{2} \sum_{i=1}^{N} (\bar{w} \cdot \overline{\phi(x_i)} - y_i)^2 + \frac{\lambda}{2} \|\bar{w}\|_2^2 \right) = 0$$

$$\frac{\partial}{\partial \bar{w}} \left(\frac{1}{2} \sum_{i=1}^{N} (\bar{w} \cdot \overline{\phi(x_i)} - y_i)^2 \right) + \frac{\partial}{\partial \bar{w}} \frac{\lambda}{2} \|\bar{w}\|_2^2 = 0$$

$$\frac{1}{2} \cdot \sum_{i=1}^{N} 2 (\bar{w} \cdot \overline{\phi(x_i)} - y_i) \overline{\phi(x_i)} + \frac{\lambda}{2} \cdot 2\bar{w} = 0$$

$$\left(\sum_{i=1}^{N} \overline{\phi(x_i)}\, \overline{\phi(x_i)}^T \right) \bar{w} - \sum_{i=1}^{N} \overline{\phi(x_i)}\, y_i - \lambda \bar{w} = 0$$

$$\phi(X)\phi(X)^T \bar{w} - \phi(X)\bar{y} - \lambda I \bar{w} = 0$$

where $\phi(X) = [\overline{\phi(x_1)}, \overline{\phi(x_2)}, \cdots \overline{\phi(x_N)}]$ ∈ (H×N) matrix

and $K = \phi(X)\phi(X)^T$ is a (m×m) (feature×feature) matrix

$K_{ij} = \phi(x_i) \cdot \phi(x_j)$

$$\therefore (\bar{K} - \lambda I)\bar{w} = \phi(X)\bar{y}$$

$$\bar{w} = [\bar{K} - \lambda I]^{-1} \phi(X)\bar{y} = (\quad)(\quad)(X^T - \lambda I)$$

$$= [\phi(X)\phi(X)^T - \lambda I]^{-1} \phi(X)\bar{y}$$

Conversion to dual form

$\phi(X)_{(h\times N)}$, $y_{(N\times 1)}$, $w_{(h\times 1)}$

Primal $= \min_{w} \frac{1}{2}\|\phi(X)^T w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$

let $u = \phi(X)^T w$ and constraint $\phi(X)^T w - y = 0$

Introduce lagrange multipliers $\alpha \in R^N$ $(N\times 1)$

$\mathcal{L}(u, \alpha, w) = \min \frac{1}{2}\|u-y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 + \alpha^T(\phi(X)^T w - u)$

$\frac{\partial \mathcal{L}}{\partial u} = 0$, we get $(u-y)^T - \alpha^T = 0 \Rightarrow u^T = y^T + \alpha^T = (y+\alpha)^T \Rightarrow$

$u = y + \alpha$

$\frac{\partial \mathcal{L}}{\partial w} = 0$, we get $\lambda w^T + \alpha^T \phi(X)^T = 0 \Rightarrow w^T = -\frac{1}{\lambda}\alpha^T \phi(X)^T \Rightarrow$

$w = -\frac{1}{\lambda}\phi(X)\alpha$

Substituting back in $\mathcal{L}(u, \alpha, w)$ to obtain the dual function $g(\alpha)$

$g(\alpha) = \frac{1}{2}\|y+\alpha-y\|_2^2 + \frac{\lambda}{2}\|-\frac{1}{\lambda}\phi(X)\alpha\|^2 + \alpha^T\left[\phi(X)^T(-\frac{1}{\lambda}\phi(X)\alpha) - y - \alpha\right]$

$g(\alpha) = \frac{1}{2}(\alpha^T\alpha) + \frac{1}{2\lambda}[\alpha^T\phi(X)^T\phi(X)\alpha] - \frac{\alpha^T}{\lambda}\phi(X)^T\phi(X)\alpha - \alpha^T y - \alpha^T\alpha$

$g(\alpha) = -\frac{1}{2}(\alpha^T\alpha) - \frac{1}{2\lambda}(\alpha^T\phi(X)^T\phi(X)\alpha) - \alpha^T y$

let $K = \phi(X)^T\phi(X)$ $(N\times N)$ be the kernel matrix

$g(\alpha) = -\alpha^T\left[\left(\frac{1}{2}I + \frac{1}{2\lambda}K\right)\alpha + y\right]$

Dual $= \max_{\alpha} -\alpha^T\left[\left(\frac{1}{2}I + \frac{1}{2\lambda}K\right)\alpha + y\right]$ K is the kernel matrix

For optimality:

$\frac{\partial}{\partial\alpha}\left[-\alpha^T\left[\left(\frac{1}{2}I + \frac{1}{2\lambda}K\right)\alpha + \alpha^T y\right]\right] = 0$

let $A = \frac{1}{2}\left(I + \frac{1}{\lambda}K\right)$

$\frac{\partial}{\partial\alpha}\left[\alpha^T A\alpha + \alpha^T y\right] = 0$

$\alpha^T(A + A^T) + y^T = 0 \Rightarrow y^T = -\alpha^T(A+A^T) \Rightarrow$

$y = -(A+A^T)\alpha = -\left[\frac{1}{2}I + \frac{1}{2\lambda}K + \frac{1}{2}I^T + \frac{1}{2\lambda}K^T\right]\alpha$

$y = -\left(I + \frac{1}{\lambda}K\right)\alpha \Rightarrow \alpha = -\left(I + \frac{1}{\lambda}K\right)^{-1} y$

Using $w = -\frac{1}{\lambda}(\phi(X)\alpha) = \frac{1}{\lambda}\phi(X)\left(I + \frac{1}{\lambda}K\right)^{-1} y$

Dual form $w = \frac{1}{\lambda}\phi(X)\left[I + \frac{1}{\lambda}\phi(X)^T\phi(X)\right]^{-1} y$

Kernel matrix

In practice, the kernel matrix is computed directly using inner dot products,

$w = \frac{1}{\lambda}X\left[I + \frac{1}{\lambda}K\right]^{-1} y$