Homework 9

Submitted By: Puneet Singhal

1) Images at end. Sorry for formatting

2)

(a) True

(b) False

(c) True

(d) A

(e) A, C

(f) Kernel SVM will be lower training error and lower test error. This is because we can choose kernels corresponding to transformations that best describe our model. This is given the Neural network in this question. If we start increasing layers in neural network and we can allow more non-linearities and hence can get better results with neural networks

3)

(a)

$x_1^2 + x_2^2$

(b) images on new page. Sorry for formatting

4)

(a)

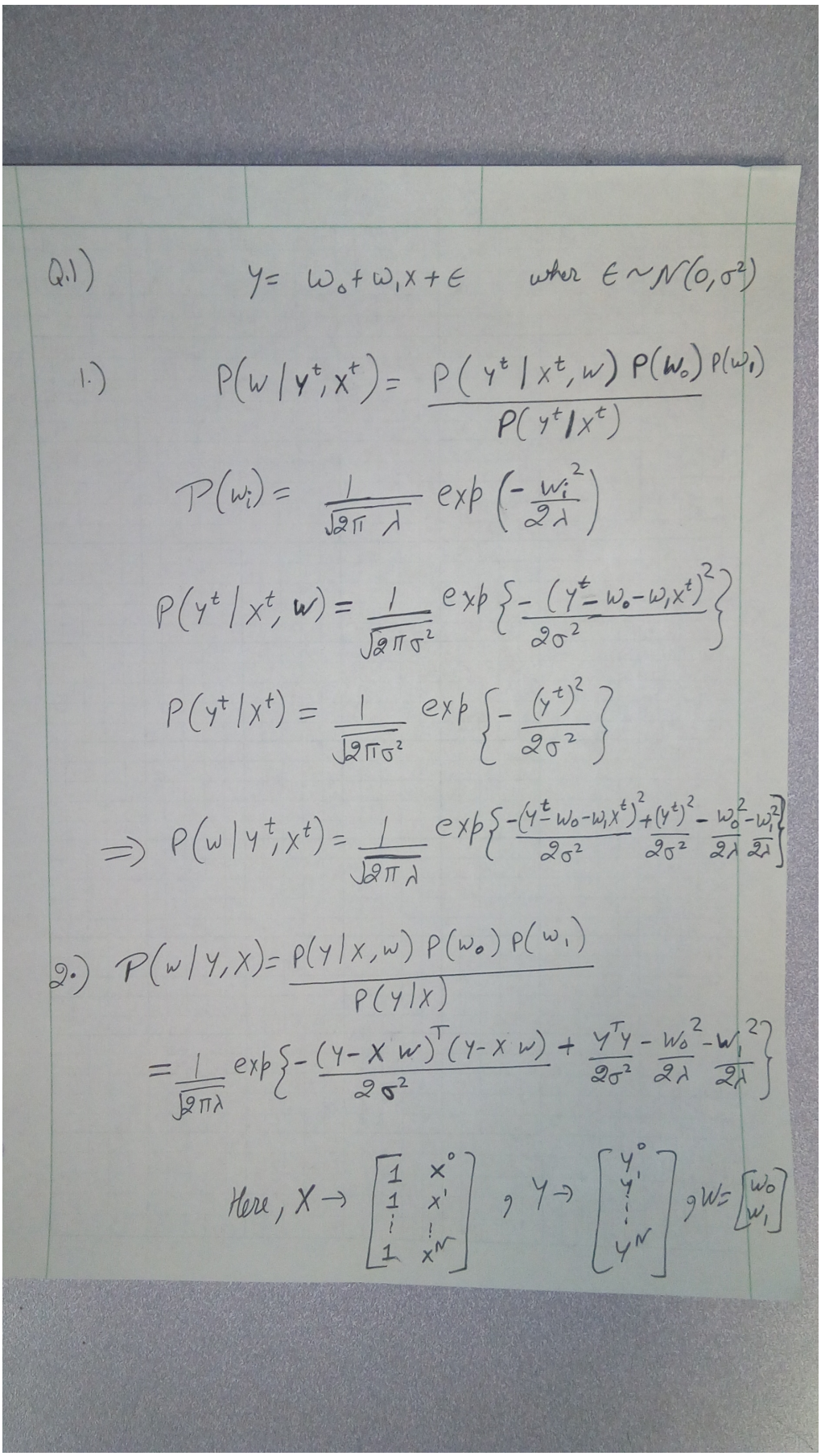
The appropriate loss function is Hinge loss as it provides tight (convex) upper bound on 0-1 loss function. While this function is convex and continuous, it is not differentiable at 1 but this can be solved using subgradient descent methods.

(b)

Mean squared error loss is appropriate as it square, absolute, continuous and differential function which will help us use techniques like gradient descent or stochastic gradient descent. This function will represent the actual cost of wrong predictions.

(c)

(a) SVM with no kernel (Soft margin): This will allow the error in classification but we will still not be able to classify 100% samples correctly.
(b) SVM with quadratic kernel with penalty on slack variable C=0: The quadratic kernel will be able to classify each sample correctly as we can see that it is possible to draw an ellipse that divides the two samples. C does not matter as there are no green sample that is surrounded by red and vice versa.
(c) SVM with quadratic kernel with penalty on slack variable C=∞: The quadratic kernel will be able to classify each sample correctly as we can see that it is possible to draw an ellipse that divides the two samples. C does not matter as there are no green sample that is surrounded by red and vice versa.
(d) Logistic regression (No kernel): Logistic regression with no kernel with have linear surface as separator which will not be able to classify all samples correctly.



Q.1) $y = w_0 + w_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$

1.) $$P(w \mid y^t, x^t) = \frac{P(y^t \mid x^t, w)\, P(w_0)\, P(w_1)}{P(y^t \mid x^t)}$$

$$P(w_i) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left(-\frac{w_i^2}{2\lambda}\right)$$

$$P(y^t \mid x^t, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y^t - w_0 - w_1 x^t)^2}{2\sigma^2}\right\}$$

$$P(y^t \mid x^t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y^t)^2}{2\sigma^2}\right\}$$

$$\Rightarrow P(w \mid y^t, x^t) = \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{-\frac{(y^t - w_0 - w_1 x^t)^2}{2\sigma^2} + \frac{(y^t)^2}{2\sigma^2} - \frac{w_0^2}{2\lambda} - \frac{w_1^2}{2\lambda}\right\}$$

2.) $$P(w \mid y, X) = \frac{P(y \mid X, w)\, P(w_0)\, P(w_1)}{P(y \mid X)}$$

$$= \frac{1}{\sqrt{2\pi\lambda}} \exp\left\{-\frac{(y - Xw)^T (y - Xw)}{2\sigma^2} + \frac{y^T y}{2\sigma^2} - \frac{w_0^2}{2\lambda} - \frac{w_1^2}{2\lambda}\right\}$$

Here, $X \to \begin{bmatrix} 1 & x^0 \\ 1 & x^1 \\ \vdots & \vdots \\ 1 & x^N \end{bmatrix}$, $y \to \begin{bmatrix} y^0 \\ y^1 \\ \vdots \\ y^N \end{bmatrix}$, $w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

5) Collaboration Policy

(a) No

(b) No

(c) No

3.) $$-\frac{(y-Xw)^T(y-Xw)}{2\sigma^2} + \frac{y^Ty}{2\sigma^2} - \frac{w^Tw}{2\lambda}$$

4.) $$\frac{d\log(P(w|y,x))}{dw} = 0$$

$$\Rightarrow -\frac{(y-Xw)^TX}{\sigma^2} - \frac{w}{\lambda} = 0$$

5.) $$\text{Gradient} = \frac{d\log(P(w|y,x)}{dw} = -\frac{(y-Xw)^TX}{\sigma^2} - \frac{w}{\lambda}$$

$$w_{new} = w_{old} + \alpha\left(-\frac{(y-X^Tw_{old})^TX}{\sigma^2} - \frac{w_{old}}{\lambda}\right)$$

6.) $P(w|y,x) \propto \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-xw)^T(y-xw)}{2\sigma^2}\right\} \frac{1}{2\lambda} \exp\left\{-\frac{|w_0|}{\lambda}\right\}$

$$\times \frac{1}{2\lambda} \exp\left\{-\frac{|w_1|}{\lambda}\right\}$$

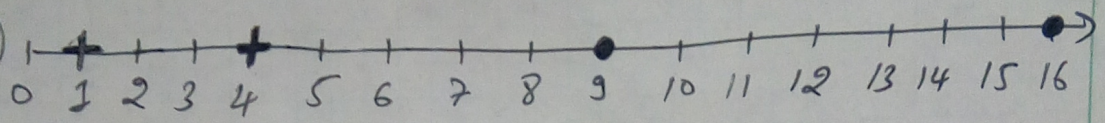
$$\log P(w|y,x) = \text{constant} - \frac{(y-xw)^T(y-xw)}{2\sigma^2} - \frac{|w_0|}{\lambda} - \frac{|w_1|}{\lambda}$$

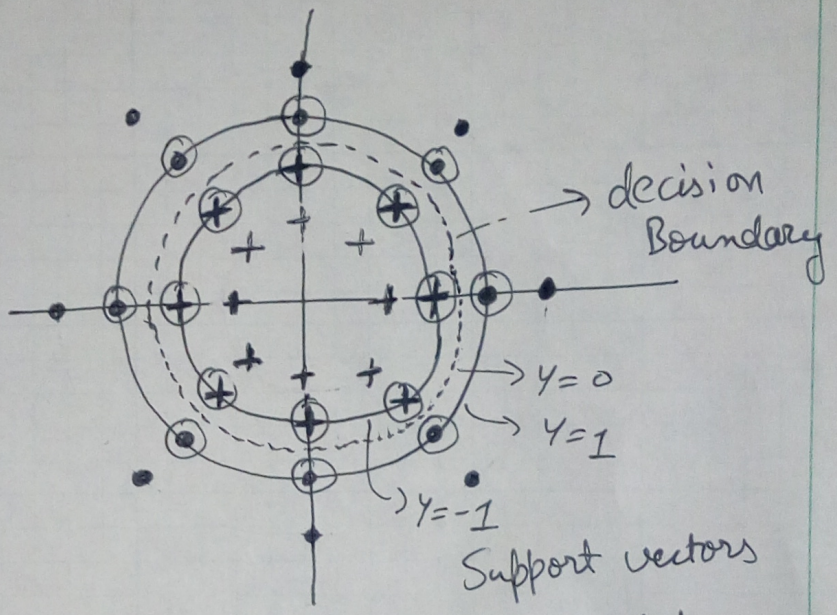
MAP: $\frac{P(w|y,x)}{dw} = 0$

$$\frac{d \log P(w|y,x)}{dw_0} = 0$$

$$\frac{d \log(P(w|y,x))}{dw_1} = 0$$

Q3)





Support vectors are circled

3.) VC dimension → 2