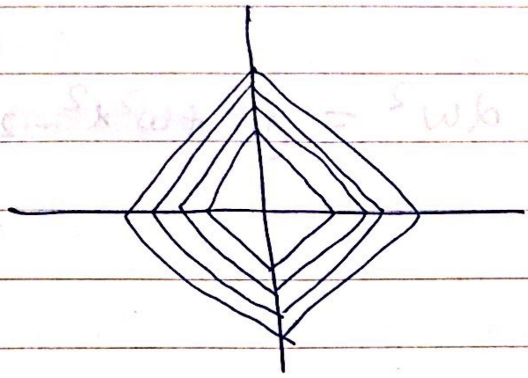
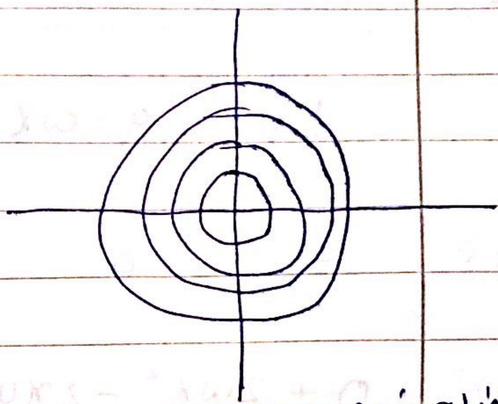
NN → ASSIGNMENT - 07 , Tutor: Redion
Priyanka Upadhyay (2581714), s8prupad@stud.uni-saarland.de
Gopal Bhattrai (7013547), gobh00001@stud.uni-saarland.de

Problem → 7.1 L1 & L2 Regularization

1.a



L_1 Regularization Norm



L_2 Regularization

1.b In L_1 Regularization, the predictor which does not have the significant impact on prediction there coefficient becomes 0 (zero) which can result as less computational cost. whereas In L_2 Regularization, coefficient never becomes equal to 0 (their value becomes close to 0). So that's why we prefer L_1 Norm over L_2.

1.c Yes. L_1 Regularization can be used for feature when the coefficient values is very small, it does not hold much significance over prediction value.
L_2 Regularization should/can be used for features with very High value of coefficient and are highly significant for outcome.

1.d No, The main aim of training model is to perform well on unseen data and Hence Regularization is done to improve Generalization / Test error by introducing some bias and decreasing the variance.

1.e with 1 Feature :

$$L_2 = (y - wx)^2 + \alpha w^2 = y^2 + w^2x^2 - 2wxy + \alpha w^2$$

Put $\frac{\partial L_2}{\partial w} = 0$

$$0 + 2wx^2 - 2xy + 2\alpha w = 0$$

$$\Rightarrow \boxed{w = \frac{xy}{x^2 + \alpha}}$$

w will become zero, when & only $\alpha \to \infty$.

1.f with 1 Feature

$$L_1 = (y - wx)^2 + \alpha w = y^2 + w^2x^2 - 2wxy + \alpha w$$

$$\frac{\partial L_1}{\partial w} = 0 \rightarrow 0 + 2wx^2 - 2xy + \alpha = 0$$

$$\boxed{w = \frac{2xy - \alpha}{2x^2}}$$

w will become , when $\alpha = 2xy$

Problem 7.2 Multi task learning and Data Augmentation

(a) Problems :

1. In some cases, Task performances decreases due to the task interference.
2. Lack of Analytical tool to predict task interference.
3. It becomes difficult to apply multi-task learning to Neural Networks, because it's feature space is given by layer-wise embeddings.

Solutions :

1. If the shared module's capacity is too large, then there is no interference between tasks.
2. Task variance is used to determine task interference.
3. Assign per-task weights for settings where different tasks share the same data but have different labels.

Problem 7.3 Early Stopping

(a) $$E = E_0 + \frac{1}{2}(w - w^*)^T H (w - w^*)$$

After τ steps, the components of the weight vector parallel to the eigen vectors of H is:

$$w_j^\tau = \{1 - (1 - \varepsilon \lambda_j)^\tau\} w_j^*$$

$$\nabla E = H\{w(\tau-1) - w^*\}$$

Scanned with CamScanner

→ So $\tau \to \infty$ i.e $\tau = 1, 2, 3, 4 \ldots$

$\tau = 1$ → $w^{(0)} = 0$ (origin)

$$w_j^{(1)} = w_i^{(0)} - \varepsilon \lambda_j \left(w_j^{(0)} - w_j^{*} \right)$$

$$= \varepsilon \lambda_j w_j^{*}$$ → computing with given value

$$= 1 - \left(1 - \varepsilon \lambda_j \right) w_j^{*}$$

$\tau = N$ → we assume the Result holds for $\tau = N - 1$

$$w_j^{N} = w_j^{N-1} - \varepsilon \lambda_j (w_j^{(N-1)} - w_j^{*})$$

$$= w_j^{(N-1)} (1 - \varepsilon \lambda_j) + \varepsilon \lambda_j w_j^{*}$$

$$= \left[1 - (1 - \varepsilon \lambda_j)^{N-1} \right] w_j^{*} (1 - \varepsilon \lambda_j)$$

$$+ \varepsilon \lambda_j w_j^{*}$$

$$= \{ (1 - \varepsilon \lambda_j) - (1 - \varepsilon \lambda_j)^{N} \} w_j^{*} + \varepsilon \lambda_j w_j^{*}$$

$$= \{ 1 - (1 - \varepsilon \lambda_j)^{N} \} w_j^{*}$$

which also Holds the Result

and provided $[1 - \varepsilon \lambda_j] < 1$

$(1 - \varepsilon \lambda_j)^{\tau} \rightarrow 0$ as $\tau \rightarrow \infty$ hence

$\{ 1 - (1 - \varepsilon \lambda_j)^{N} \} \rightarrow 1$ and we

can prove that $w^{(\tau)} \rightarrow w^{*}$

(b) If τ is finite but $\lambda_j \gg (\varepsilon \tau)^{-1}$ then τ must still be large, since $\lambda_j \varepsilon \tau \gg 1$, even though $|1 - \varepsilon \lambda_j| < 1$

if τ is lager than from above we can show that $w_j^{(\tau)} \simeq w_j^{*}$

if $\lambda_j \ll (\varepsilon \tau)^{-1}$, then it means $\varepsilon \lambda_j$ must be small, since $\varepsilon \lambda_j \tau \ll 1$ & τ is an integer greater than or equal to 1

$$w_j^{(\tau)} \ll |w_j^{*}|$$

(c) we know that, when regularization parameter (α) is much larger than one of the eigen values (λ_j) then corrosponding parameter value w_i will close to 0.

In the same way when α is smaller than λ_j, w_i, will be close to its maximum likelihood value.

Thus we can say that α is analogous to $\epsilon\tau$

Problem 7.4 Universal Approximation Theorem

a) This paper contribute and validate the Idea that Multilayer Feedforward Architecture provides Neural Network the potential of being universal learning machines and not the choice of Activation Function.

b) The main idea of the paper is that activation function is more important for approximation rather than the depth of the Network. It states that Neural Networks with a single hidden layer can accurately make predictions when provided with a continuous sigmoidal Non-Linear activation function.

c) A Deeper Network with multiple layers effectively learns various features at different levels of abstraction and hence provide better generalization.