Part3

3. Layer-wise Analysis and Design of Deep Neural Networks

3.1 Two Data Complexity measures

- Separation index (SI)
- Smoothness index(SmI)

3.2 Layer-wise Analysis by Separation and Smoothness indices

- Dataset evaluation, ranking and dividing (SI, SmI)
- Subset Selection (SI, SmI)
- Layer-wise Model evaluation (SI, SmI)
- Pre-train Model ranking (SI, SmI)
- Model Confidence and Guarantee (SI, SmI)

3.3 Layer-wise Design by Separation and Smoothness indices

- Model Compressing(SI, SmI)
- Forward learning in the first layer(SI, SmI)
- Layer-wise forward learning(SI, SmI)
- Layer-wise branching(SI, SmI)
- Layer-wise Fusion(SI, SmI)
- Forward Design(SI, SmI)
- Forward Multi-Task Design(SI, SmI)

3.4 Related works in local Layer-wise learning

Indicator	Research (state)
SI	Initial studies have been done
SmI	Initial studies have been done
SI	There are some prepared/under-review works
SmI	There are some prepared/under-review works
SI	New idea
SmI	New idea

3.1 Two Data Complexity measures

3.1.1 Separation index

- First order SI
- High order SI
- High order soft SI

3.1.2 Smoothness index

- First order SmI
- High order SmI
- High order soft SmI

Data Complexity measures

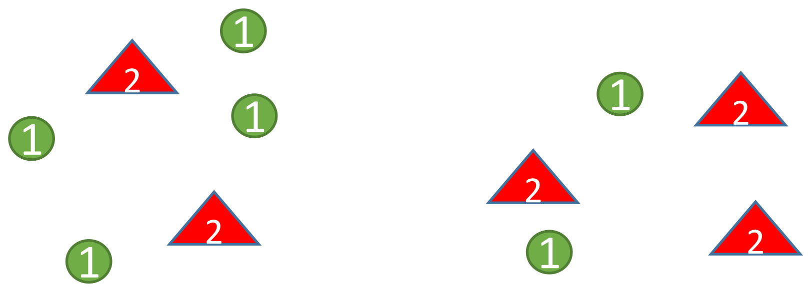
Complexity measures	Overall evaluating approach
√ Feature-based	Discovering informative features by evaluating each feature independently (Orriols-Puig et al., 2010; Cummins, 2013))
Linearity separation	Evaluating the linearly separation of different classes (Bottou & Lin, 2007)
√ Neighborhood	Evaluating the shape of the decision boundary to distinguish different classes overlap (Lorena et al., 2012; Leyva et al., 2014)
√ Network	Evaluating the data dataset structure and relationships by representing it as a graph (Garcia et al., 2015)
√ Dimensionality	Evaluating the sparsity of the data and the average number of features at each dimension (Lorena et al., 2012; Basu & Ho, 2006)
√ Class imbalanced	Evaluating the proportion of dataset number between different classes (Lorena et al., 2012)

Table 1. Some complexity measures and their evaluating approaches in a classification problem

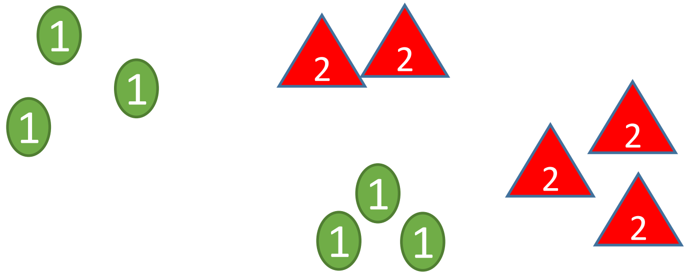
Two Complexity measures

1. A separation measure (in classification problems)

   It shows that how much input data points separate the labels from each others.



(Less separated)

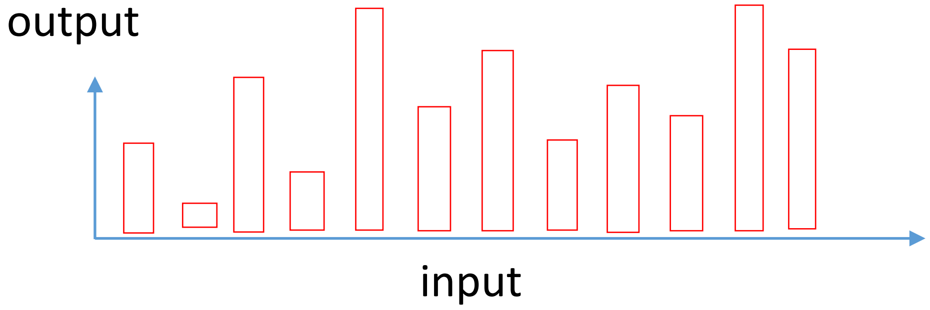


(More separated)

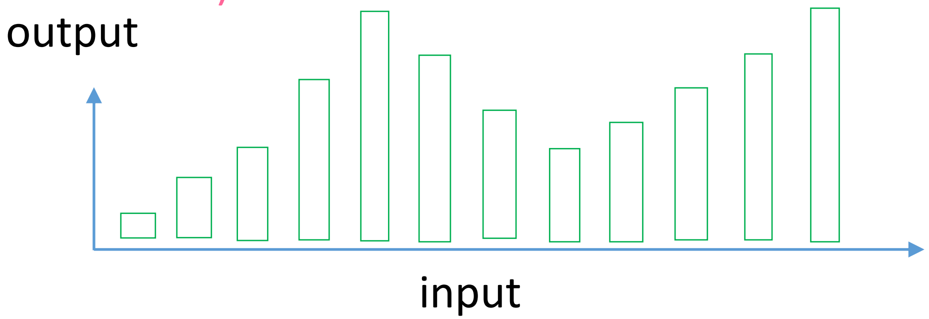
2. An smoothness measure (in regression problems)

   It shows that how much input data points make the output targets smooth

(Less Smooth)



(More Smooth)



Separation index (SI)

“SI” measures that how much input data points separate class labels from each others.

3.1.1 Separation index (SI)

1. First order SI

$Data = \{(\boldsymbol{x}^i, l^{\mathrm{i}})\}_{i=1}^{m}$ $\forall i$: $\boldsymbol{x}^i \epsilon R^{n\times 1}$ $l^{\mathrm{i}} \epsilon \{1,2,\ldots,n_C\}$ n_C:number of classes

*it is assumed that "Data" is a measured sample from a domain with high enough diversity.

*$\boldsymbol{x}^i$ may have any format (video, image, time series, etc.) ; however, to compute SI, it must be reshaped as a vector.

$$\mathrm{SI}(Data) = \frac{1}{m}\sum_{q=1}^{m} \delta\left(l^i, l^{i^*}\right)$$

$$i^* = \underset{\forall q \neq i}{\arg\ min} \left\| \boldsymbol{x}^i - \boldsymbol{x}^q \right\| \qquad \delta\left(l^i, l^{i^*}\right) = \begin{cases} 1 & l^i = l^{i^*} \\ 0 & else \end{cases}$$

kronecker delta

*$\|\cdot\|$ denotes Euclidian distance (L_2 norm) but it may be another distance definition such as Lp norm: $\left\| \boldsymbol{x}^i - \boldsymbol{x}^j \right\|_{L_p} = \sqrt[p]{\sum_{k=1}^{n} |\boldsymbol{x}^i(k) - \boldsymbol{x}^j(k)|^p}$

** It is assumed that the input data is normalized at each dimension just before computing separation index.

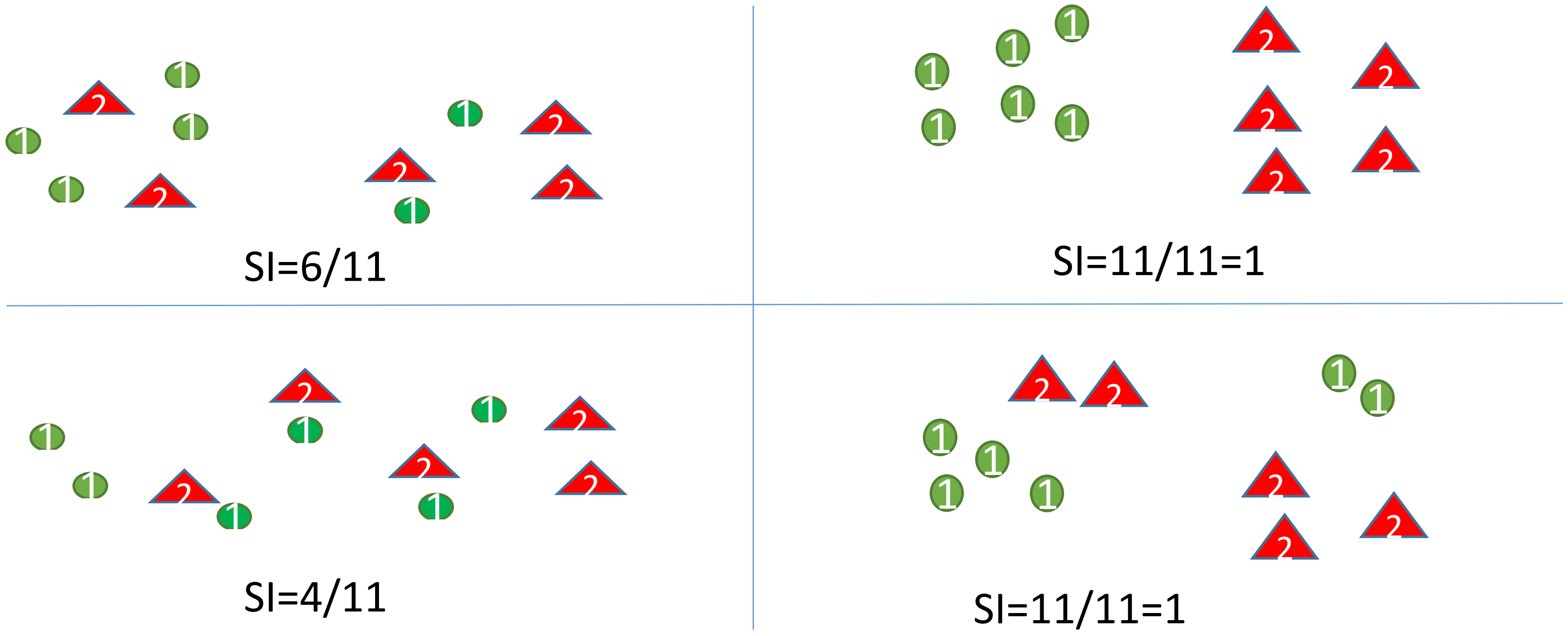
Some notes

1. "SI" is a normalized index between zero and one: $\mathrm{SmI} \in [0,1]$
2. $SI \to 1$ $(Sepration\ is\ maximmum)$ and $SI \to 0 (Sepration\ is\ minimmum)$
3. "SI" counts (average of) all data points whose nearest neighbors have the same label
4. "SI" is equal to the accuracy of the nearest neighbor classifier as a non-parametric model. Hence, SI is an informative index having strong correlation with the best accuracy one can access by a model without filter process.
5. SI does not change against shift and scales of data points.

$$\forall \beta \neq 0, \forall \alpha \neq 0, \forall x_0, \forall l_0 \qquad SI(\{(\boldsymbol{x}^i, l^i)\}_{i=1}^m) = SI(\{(\beta \boldsymbol{x}^i + \boldsymbol{x}_0, \alpha l^i + l_0)\}_{i=1}^m)$$

6. Separatin index of $the\ target\ labels\ with\ themselves\ is\ maximum$: $SI(\{(l^i, l^i)\}_{i=1}^m)$=1

Two dimensional examples (binary classification)



Some notes

- To have a high SI, It is enough that examples of each class become near and near together in some regions
- The number of regions is not important but each region must have at least two members.
- The shape of each region is not important.

The distance matrix

- To achieve SI, matrix distance of all data points must be computed (to get nearest neighbor for each data point)

$$Data = \{(\boldsymbol{x}^i, l^{\mathrm{i}})\}_{i=1}^{m} \qquad \boldsymbol{x}^i \epsilon R^{n\times 1}$$

Distance matrix: $D = [d_{ij}] \quad d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$

Steps

1- Provide data Matrix:$\mathrm{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m]^T, \quad \mathrm{X}\epsilon R^{m\times n}$

2- M= $\mathrm{XX^T}, \quad \mathrm{M}\epsilon R^{m\times m}$

3- $\mathrm{d} = \mathrm{diag(M)}, \quad \mathrm{d}\epsilon R^{m\times 1}$

4- W=[d,d,…,d], $\quad \mathrm{W}\epsilon R^{m\times m}$

5- Distance matrix is computed as follows:

$$\mathrm{D} = \mathrm{W} + \mathrm{W^T} - 2\mathrm{M}$$

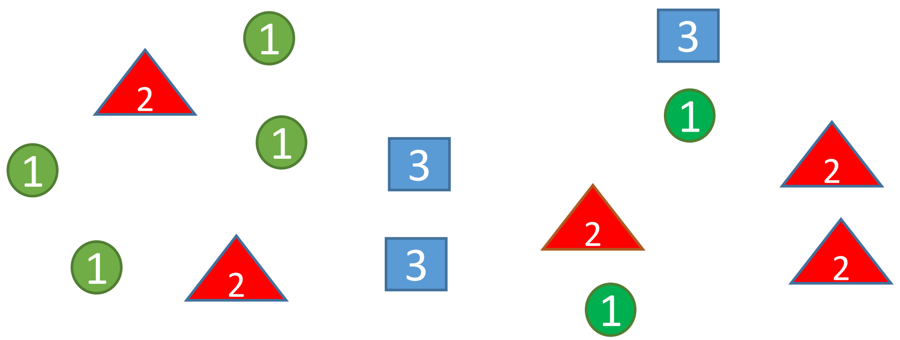
Separation index for Each Class

$$\mathrm{SI}_c(Data) = \frac{1}{m_c}\sum_i \delta(l^i, c)\delta(l^i, l^{i^*}) \qquad c=1,2,..,n_C$$

m_c=$\sum_i \delta(l^i, c)$ $\qquad$ m_c: number of all data points x^i which $l^i = c$

$$SI(Data) = \frac{1}{m}\sum_{c=1}^{n_C} m_c SI_c(Data) \qquad \sum_{c=1}^{n_C} m_c = m$$

A two dimensional illustrative example



$n_C = 3, \qquad c = 1,2,3$

$SI_1(Data) = 4/6$
$SI_2(Data) = 2/5$
$SI_3(Data) = 2/3$
SI=(4+2+2)/(6+5+3)=8/14

* For when for each class c: $m_c = \frac{m}{n_C}$ and a sufficient high number of data points are distributed with a *uniformly distributed random* variable then it is expected that $SI \to 1/n_C$

2. High order SI

$Data = \{(\boldsymbol{x}^i, l^{\mathrm{i}})\}_{i=1}^{m}$ $\forall i$: $\boldsymbol{x}^i \epsilon R^{n\times 1}$ $l^{\mathrm{i}} \epsilon \{1,2,\ldots,n_C\}$ n_C:number of classes

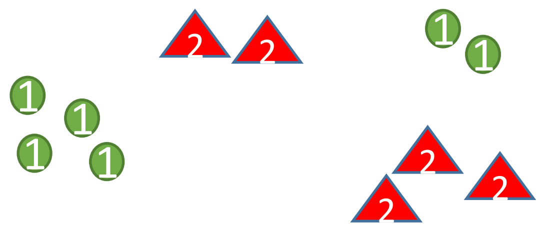
$$\mathrm{SI}^{\mathrm{r}}(Data) = \frac{1}{m}\sum_{q=1}^{m} \delta_{hard}\left(l^i, l^{i_1^*}, \cdots, l^{i_r^*}\right)$$ r: the order of "SI"

$$i_j^* = \underset{\forall q \neq i, i_1^*, \cdots, i_{j-1}^*}{\arg} \; min\left\|\boldsymbol{x}^i - \boldsymbol{x}^q\right\| \qquad \delta_{hard}\left(l^i, l^{i_1^*}, \cdots, l^{i_r^*}\right) = \prod_{j=1}^{r} \delta\left(l^i, l^{i_j^*}\right)$$

- $\mathrm{SI}^{\mathrm{r}} \in [0,1]$
- "SI^{r}" counts (average of) all data points whose all "r" nearest neighbors have the same label
- SI^{r} considers more restricted condition of separation than SI^{j} $(j < r)$.
- For each "Data" we have: $\mathrm{SI}^{\mathrm{r}} \leq \mathrm{SI}^{\mathrm{r}-1} \leq \cdots \leq \mathrm{SI}^{1}$ $\qquad \mathrm{SI}^{1} = \mathrm{SI}$

Two illustrative Examples

SI^1 =11/11
SI^2 =11/11
SI^3 =11/11
SI^4 =11/11



SI^1 =11/11
SI^2 =7/11
SI^3 =4/11
SI^4 =0

3. High order soft SI

$Data = \{(\boldsymbol{x}^i, l^{\mathrm{i}})\}_{i=1}^{m}$ $\forall i$: $\boldsymbol{x}^i \epsilon R^{n\times 1}$ $l^{\mathrm{i}} \epsilon \{1,2,\ldots,n_C\}$ n_C:number of classes

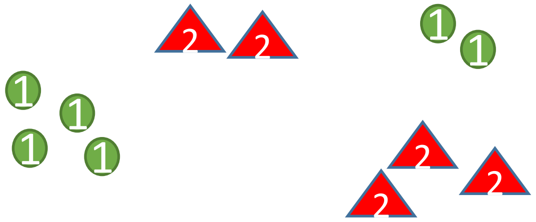
$$\mathrm{SI}^{\mathrm{r}}_{\mathrm{soft}}(Data) = \frac{1}{m}\sum_{q=1}^{m} \delta_{soft}\left(l^i, l^{i_1^*}, \cdots, l^{i_r^*}\right)$$ r: the order of "SI"

$$i_j^* = \underset{\forall q\neq i, i_1^*, \cdots, i_{j-1}^*}{\arg} \quad min\left\|x^i - x^q\right\| \qquad \delta_{soft}\left(l^i, l^{i_1^*}, \cdots, l^{i_r^*}\right) = \sum_{j=1}^{r} \delta\left(l^i, l^{i_j^*}\right)/r$$

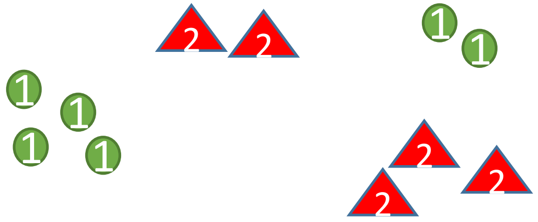
- $\mathrm{SI}^{\mathrm{r}}_{\mathrm{soft}} \in [0,1]$
- $\mathrm{SI}^{\mathrm{r}}_{\mathrm{soft}}$ considers less restricted condition of separation than SI^{r}

$\mathrm{SI}^{\mathrm{r}}_{\mathrm{soft}} \geq \mathrm{SI}^{\mathrm{r}}$ and $\mathrm{SI}^{1}_{\mathrm{soft}} = \mathrm{SI}^{1}$

Two illustrative Examples



SI^1 =11/11
SI^2 =7/11
SI^3 =4/11
SI^4 =0



SI^1_{soft} =11/11
SI^2_{soft}=(4+3+0.5+0.5)/11=8/11
SI^3_{soft}=(4+3(2/3)+4*(1/3)/11=8.33/11
SI^4_{soft}=(4*(3/4)+2*(1/4)+2*(1/4)+3*(2/4))/11
=6.5/11

Smoothens index (SmI)

SmI measures how much input data points make the output targets smooth

3.1.2 Smoothness index (SI)

A smoothness measure for regression problem

1. First order SI

$Data = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{i=1}^{m}$ $\forall i$: $\boldsymbol{x}^i \epsilon R^{n\times 1}$, $\boldsymbol{y}^i \epsilon R^{o\times 1}$ o :number of outputs

*it is assumed that Data is a measured sample with high enough diversity.

*x^i and y^i may have any format (video, image, time series, etc.) ; however, to compute SmI, it must be reshaped as a vector.

$$\mathrm{SmI}(Data) = \frac{1}{m}\sum_{q=1}^{m} \frac{\|\boldsymbol{y}^{imax} - \boldsymbol{y}^{i*}\|}{\|\boldsymbol{y}^{imax} - \boldsymbol{y}^{imin}\|}$$

$$i^* = \underset{\forall q\neq i}{\arg\, min}\|\boldsymbol{x}^i - \boldsymbol{x}^q\| \qquad imax = \underset{\forall q}{\arg\, max}\|\boldsymbol{y}^i - \boldsymbol{y}^q\| \qquad imin = \underset{\forall q\neq i}{\arg\, min}\|\boldsymbol{y}^i - \boldsymbol{y}^q\|$$

*$\|\cdot\|$ denotes Euclidian distance (L_2 norm) but it may be another distance definition such as L_p norm.

** It is assumed that the input and target output data are normalized at each dimension just before computing the smoothness index.

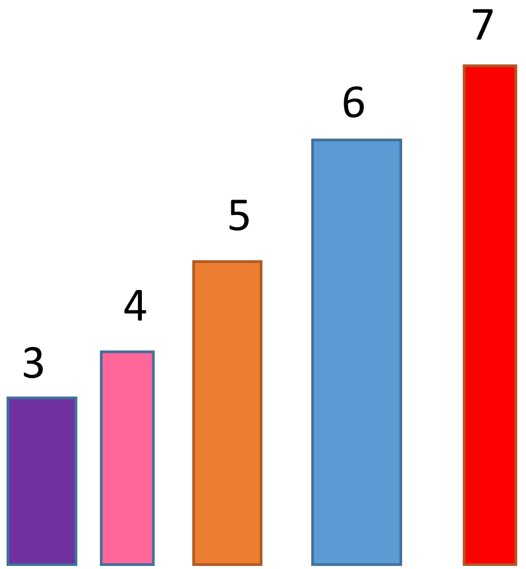
Some notes

1. “SmI” is a normalized index between zero and one: $\mathrm{SmI} \in [0,1]$
2. $SmI \to 1$ $(Smoothness\ is\ maximmum)$ and $SmI \to 0$ $(Smoothness\ is\ minimmum)$
3. “SmI” measures that how nearness of input data leads to nearness of target data.
4. Assuming, the target outputs are outputs of a classification problem in “one-hot” format, SmI is actually measure the separation index: $\mathrm{SmI} = \mathrm{SI}$
5. Increasing the number of classes and considering a nearness among every two classes, SI is interpreted as a smoothness index.
6. SmI does not change for arbitrary position shift and (scalar) scale of the data

$\forall\beta \neq 0, \forall\alpha \neq 0, \forall x_0, \forall y_0 \qquad \mathrm{SmI}(\{(x^i, y^i)\}_{i=1}^{m}) = SmI(\{(\beta x^i + x_0, \alpha y^i + y_0)\}_{i=1}^{m})$

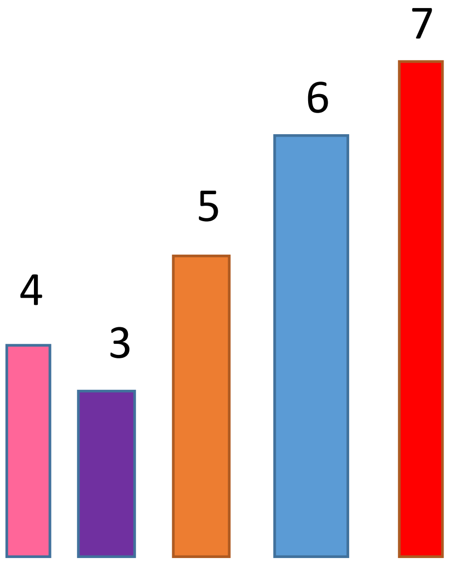
7. Smoothness index of target outputs $with\ themselves\ is\ aximum$: $SmI(\{(y^i, y^i)\}_{i=1}^{m})$=1

One-dimensional illustrative examples



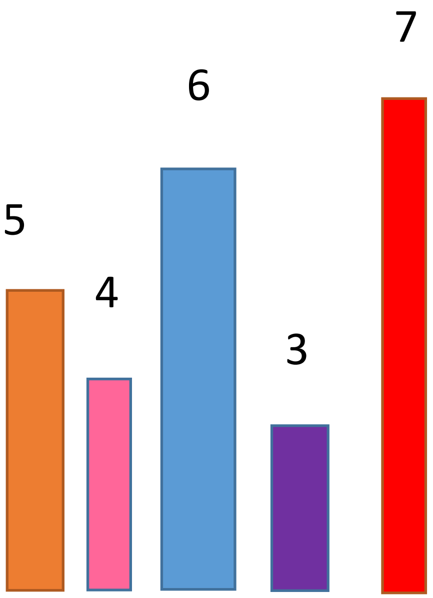
$$SmI = \frac{1}{5}\left(\frac{7-4}{7-4}+\frac{7-3}{7-3}+\frac{7-4}{7-4}+\frac{5-3}{5-3}+\frac{6-3}{6-3}\right)$$

$$SmI = 1$$



$$SmI = \frac{1}{5}\left(\frac{7-4}{7-4}+\frac{7-3}{7-3}+\frac{7-4}{7-3}+\frac{5-3}{5-3}+\frac{6-3}{6-3}\right)$$

SmI=0.95



$$SmI = \frac{1}{5}\left(\frac{7-4}{7-4}+\frac{7-5}{7-3}+\frac{4-3}{5-3}+\frac{7-6}{7-4}+\frac{3-3}{6-3}\right)$$

SmI=.466

2. High order SmI

$Data = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{i=1}^{m} \quad \forall i:\ \boldsymbol{x}^i \epsilon R^{n\times 1} \quad \boldsymbol{y}^{\mathrm{i}} \epsilon R^{o\times 1}$

$$\mathrm{SmI}^{\mathrm{r}}(Data) = \frac{1}{m}\sum_{q=1}^{m} \min\left\{\frac{\left\|\boldsymbol{y}^{imax} - \boldsymbol{y}^{i_j^*}\right\|}{\left\|\boldsymbol{y}^{imax} - \boldsymbol{y}^{imin_j}\right\|}\right\}_{j=1}^{r}$$

r: the order of "SmI"

$$i_j^* = \underset{\forall q\neq i, i_1^*, \cdots, i_{j-1}^*}{\arg \min} \|\boldsymbol{x}^i - \boldsymbol{x}^q\| \qquad imin_j = \underset{\forall q\neq i, imin_1, \cdots, imin_{j-1}}{\arg \min} \|\boldsymbol{y}^i - \boldsymbol{y}^q\|$$

- $\mathrm{SmI}^{\mathrm{r}} \in [0,1]$
- $\mathrm{SmI}^{\mathrm{r}}$ considers more restricted condition of smoothness than $\mathrm{SmI}^{\mathrm{j}}$ $(j < r)$.
- For each "Data" we have: $\mathrm{SmI}^{\mathrm{r}} \leq \mathrm{SmI}^{\mathrm{r}-1} \leq \cdots \leq \mathrm{SmI}^{1}$ $\qquad$ SmI^{1}=SmI

2. High order soft SmI

$Data = \{(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{i=1}^{m} \quad \forall i:\ x^i \epsilon R^{n\times 1} \quad y^{\mathrm{i}} \epsilon R^{o\times 1}$

$$\mathrm{SmI}_{\mathrm{soft}}^{\mathrm{r}}(Data) = \frac{1}{m}\sum_{q=1}^{m} \frac{\left\| \boldsymbol{y}^{imax} - \underset{j}{\mathrm{mean}}\, \boldsymbol{y}^{i_j^*} \right\|}{\left\| \boldsymbol{y}^{imax} - \underset{j}{\mathrm{mean}}\, \boldsymbol{y}^{imin_j} \right\|} \quad \mathrm{j} = 1,2,\ldots,\mathrm{r} \quad \mathrm{r}: \text{the order of "SmI"}$$

$$i_j^* = \underset{\forall q \neq i, i_1^*, \cdots, i_{j-1}^*}{\arg\ min} \|\boldsymbol{x}^i - \boldsymbol{x}^q\| \qquad imin_j = \underset{\forall q \neq i, imin_1, \cdots, imin_{j-1}}{\arg\ min} \|\boldsymbol{y}^i - \boldsymbol{y}^q\|$$

- $\mathrm{SmI}_{\mathrm{soft}}^{\mathrm{r}} \in [0,1]$
- $\mathrm{SmI}_{\mathrm{soft}}^{\mathrm{r}}$ considers less restricted condition of smoothness than $\mathrm{SmI}^{\mathrm{r}}$

$$\mathrm{SmI}_{\mathrm{soft}}^{\mathrm{r}} \geq \mathrm{SmI}^{\mathrm{r}} \quad \text{and} \quad \mathrm{SmI}_{\mathrm{soft}}^{1} = \mathrm{SmI}^{1}$$

3.2 Analysis by Separation and Smoothness indices

3.2.1 Dataset evaluation, ranking and dividing

3.2.2 Subset Selection

3.2.3 Layer-wise Model evaluation

3.2.4 Pre-train Model ranking

3.2.5 Model Confidence and Guarantee