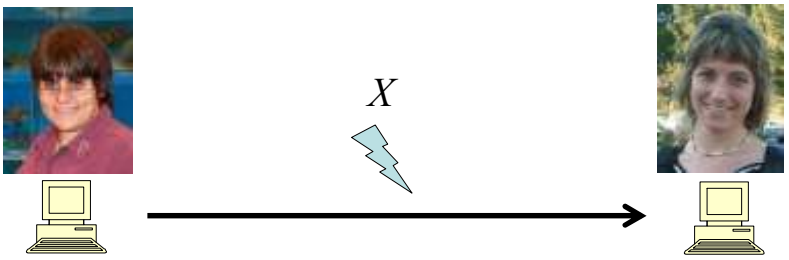
An Information Theoretic Perspective on Multiple Classifier Systems

Gavin Brown
University of Manchester, UK
gavin.brown@manchester.ac.uk

A Communications Channel



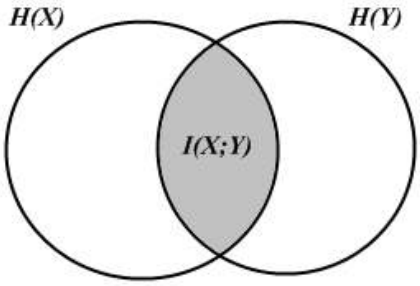
$X = encode(Y)$ $\qquad\qquad$ $\hat{Y} = decode(X)$

X = features, *Y* = class label, *decode(.)* = predictor

Mutual Information

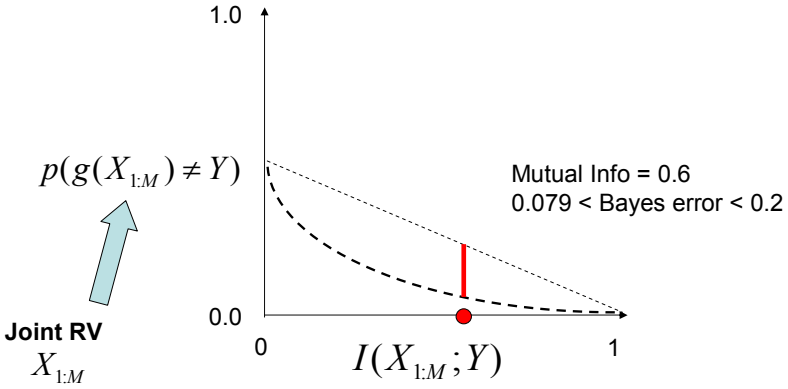
$I(X;Y)$

- Dependence measure between random variables.
- Zero when X,Y are statistically independent.
- Increases to a max when simple relationship.



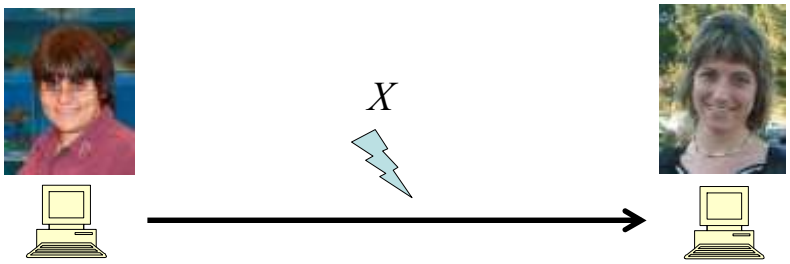
H(X) and H(Y) are entropies.
I(X;Y) is the SHARED information.

Mutual Information bounds the Bayes Rate



CONCLUSION?
Maximise mutual information between features and labels!
If high mutual info, there exists a simple function to predict Y.

A Communications Channel



$X = encode(Y)$ $\quad\quad$ $\hat{Y} = decode(X)$

X = features, *Y* = class label, *decode(.)* = predictor

X = *classifiers,* *Y* = class label, *decode(.)* = combiner

Multiple Classifier Systems

CONCLUSION?
Maximise mutual information between CLASSIFIERS and labels!
If high mutual info, there exists a simple COMBINER to predict Y.

Select or build our classifier set $S = \{X_1, ..., X_M\}$

such that we maximise $I(X_{1:M}; Y)$

So can we get new insights from this?

The most quoted phrase in our community : *"higher diversity"* …

$$ensemble\ performance = accuracy + diversity$$

Why are/were we so obsessed?

$$\text{MSE} = \text{bias}^2 + \text{variance}$$

$$E_{\text{add}}^{\text{ave}} = E_{\text{add}}\left(\frac{1+\rho(L-1)}{L}\right)$$

normalize

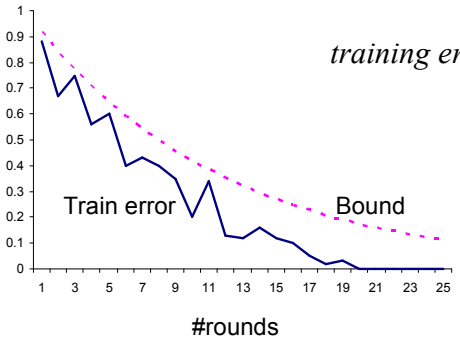
Plus "Bias+Variance+Covariance" extension for linear combiners
(basis of Tumer+Ghosh 1996)

Depends on loss function & combiner! What about majority vote?

$$majority\ vote\ error = \ ? \ + \ ?$$

Not possible – many-one-mapping, loss of information

Adaboost does not try to minimize training error



$$training\ error \le \prod_t \left[2\sqrt{\varepsilon_t(1-\varepsilon_t)}\right]$$

Adaboost minimizes a BOUND on the error.
Just one of many examples of using a **surrogate** loss function.

So, we will decompose a BOUND on the error – mutual information…

Mutual Information

$$I(X_1;X_2)=\sum_{x_1\in X_1}\sum_{x_2\in X_2}p(x_1x_2)\log\frac{p(x_1x_2)}{p(x_1)p(x_2)}$$

Conditional Mutual Information

$$I(X_1;X_2\mid Y)=\sum_{y\in Y}p(y)\sum_{x_1\in X_1}\sum_{x_2\in X_2}p(x_1x_2\mid y)\log\frac{p(x_1x_2\mid y)}{p(x_1\mid y)p(x_2\mid y)}$$

Multivariate Mutual Information

- Shannon (1948) defined the field.
- McGill (1954) proposed a multivariate extension.

$$I(\{X_1,X_2,X_3\})=I(X_1;X_2\mid X_3)-I(X_1;X_2)$$

The **difference** in dependence, **before** and **after** observing X_3.

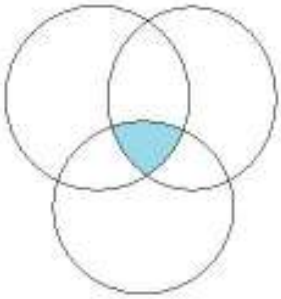
Recursive definition :

$$I(\{X_1,X_2,X_3,X_4\})=I(\{X_1,X_2,X_3\}\mid X_4)-I(\{X_1,X_2,X_3\})$$

Expanding the Mutual Information

The information can be expanded ….

$$I(X_{1:M};Y) = \sum_{i=1}^{M} I(X_i;Y) - \sum_{X\subseteq S} I(\{X\}) + \sum_{X\subseteq S} I(\{X\}\,|\,Y)$$



…ıcy”

…ation,

redundancy

“redundancy”

conditional “redundancy”

“**Multivariate**” mutual information (McGill, 1954)

Extension of Shannon’s 1948 theory. Set argument, reduces to Shannon for set size 2

But, it’s not that straightforward….

$$I(X_{1:M};Y) = \sum_{i=1}^{M} I(X_i;Y) - \sum_{\substack{X\subseteq S\\|X|=2}} I(\{X\}) + \sum_{\substack{X\subseteq S\\|X|=2}} I(\{X\}\,|\,Y)$$
$$- \sum_{\substack{X\subseteq S\\|X|=3}} I(\{X\}) + \sum_{\substack{X\subseteq S\\|X|=3}} I(\{X\}\,|\,Y)$$
$$\vdots$$
$$- \sum_{\substack{X\subseteq S\\|X|=M}} I(\{X\}) + \sum_{\substack{X\subseteq S\\|X|=M}} I(\{X\}\,|\,Y)$$

+ 3-way diversity
+ …-way diversity
| M-way diversity

| 3-way diversity
| …-way diversity
+ M-way diversity

High Order Diversity!

$$I(X_{1:M};Y) = \textit{Individual Mutual Info} + \textit{2-way diversity (pairwise)} + \textit{3-way diversity} + \textit{...-way diversity} + \textit{M-way diversity}$$

What does this tell us?

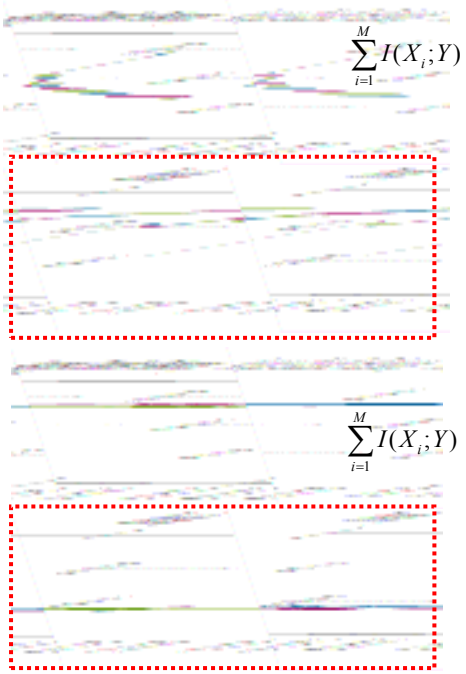
- Pairwise diversity measures are not enough…
- But calculating high order components is intractable!
  (M-dimensional probability distributions)

Monitoring Low-Order Diversity Components

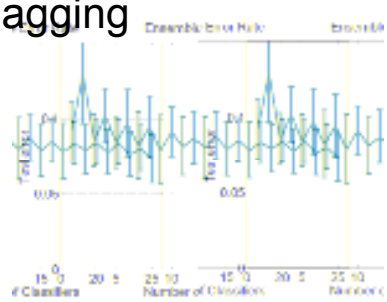
$$I(X_{1:M};Y) \approx \underbrace{\sum_{i=1}^{M} I(X_i;Y)}_{\text{“relevancy”}} - \underbrace{\sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_i;X_j) + \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_i;X_j \mid Y)}_{\text{“diversity”}}$$

$$I(X_{1:M};Y) = \textit{Individual Mutual Info} + \textit{2-way diversity (pairwise)} + \sout{\textit{3-way diversity}} + \sout{\textit{... way diversity}} + \sout{\textit{M-way diversity}}$$

Adaboost

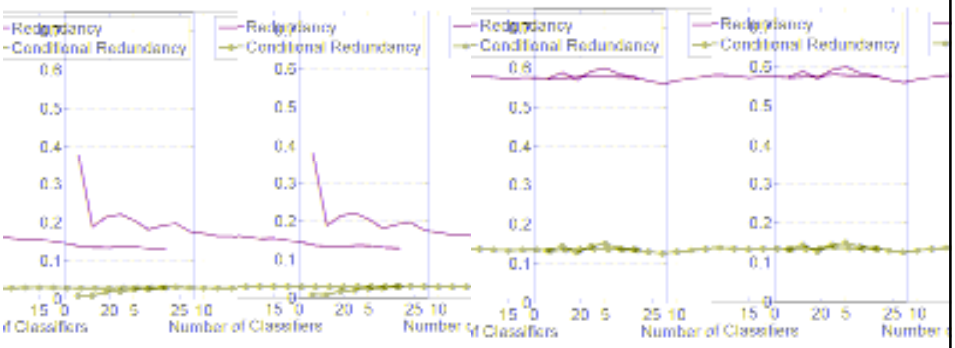


Bagging



Adaboost “Diversity”

Bagging “Diversity”



$$I(X_{1:M};Y) \approx relevancy - redundancy + condredundancy$$

$$I(X_{1:M};Y) \approx \sum_{i=1}^{M} I(X_i;Y) - \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_i;X_j) + \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_i;X_j \mid Y)$$

Conclusions

- Information Theory provides a neat way of thinking about MCS
- Mutual Information reveals **natural** measures of diversity
- Diversity exists on **multiple** levels! High and low orders.

FUTURE WORK

1. Characterize the expansion terms
   - what algorithms generate high order diversity?
   - how can we control it?
2. Understanding semi-sup / generative ensembles? Manuela…..

$$I(X_{1:M};Y) \approx \sum_{i=1}^{M} I(X_i;Y) - \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_i;X_j) + \sum_{j=1}^{M}\sum_{k=j+1}^{M} I(X_i;X_j \mid Y)$$

FIN

Mutual Information

$$I(X_1;X_2)=\sum_{x_1\in X_1}\sum_{x_2\in X_2}p(x_1x_2)\log\frac{p(x_1x_2)}{p(x_1)p(x_2)}$$

Conditional Mutual Information

$$I(X_1;X_2\mid Y)=\sum_{y\in Y}p(y)\sum_{x_1\in X_1}\sum_{x_2\in X_2}p(x_1x_2\mid y)\log\frac{p(x_1x_2\mid y)}{p(x_1\mid y)p(x_2\mid y)}$$

Multivariate Mutual Information

- Shannon (1948) defined the field.
- McGill (1954) proposed a multi-variate extension.

$$I(\{X_1,X_2,X_3\})=I(X_1;X_2\mid X_3)-I(X_1;X_2)$$

The **difference** in dependence, **before** and **after** observing X_3.

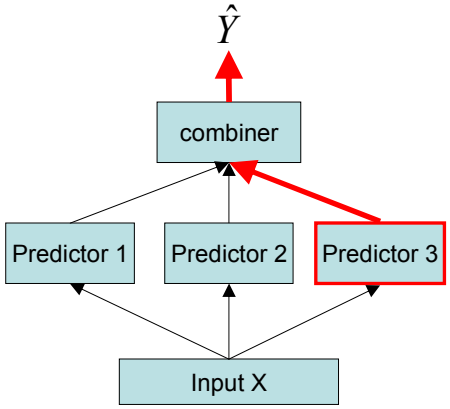
Recursive definition :

$$I(\{X_1,X_2,X_3,X_4\})=I(\{X_1,X_2,X_3\}\mid X_4)-I(\{X_1,X_2,X_3\})$$

Ensemble Diversity…

Biggest buzzword in the field…. but what **is** it?!

Fundamentally, **DIVERSITY** is a **CREDIT ASSIGNMENT** problem.



*Given that the ensemble makes an error, how "**responsible**" is predictor 3 ?*

We have to be able to go back "through" the combiner.