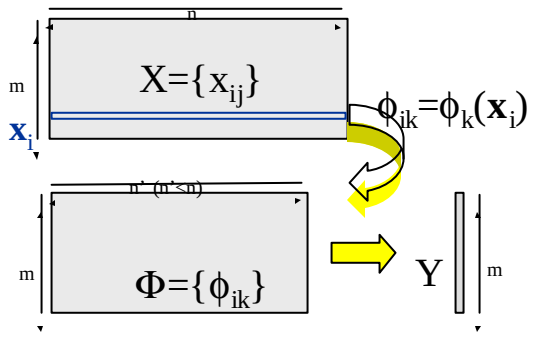
Lecture 11: Information Theoretic Methods

Isabelle Guyon
guyoni@inf.ethz.ch

Book Chapter 6 and
http://www.jmlr.org/papers/volume3/torkkola03a/torkkola03a.pdf

Mutual Information as "Information Gain"

Feature Transforms



Information, Entropy

- For an event happening with probability p, the quantity of information is –log p.
- If the log is of base 2, the unit is a "bit".
- The average quantity of information over all events is the entropy:

$$H = - \Sigma\ p \log p$$

Conditional Entropy

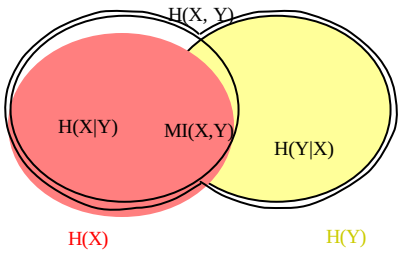
- Our case: Y=class labels, Φ feature representation.
- Entropy: $H(Y) = -\sum_y p(y) \log p(y)$
- Conditional entropy:

$$H(Y|\Phi) = -\int_f p(f) \sum_y p(y|f) \log p(y|f)\, df$$

Mutual Information

- Information Gain: The amount of class uncertainty reduction by observing Φ is:
- $MI(Y, \Phi) = H(Y) - H(Y|\Phi)$
- $MI(Y, \Phi) = \sum_y \int_f p(y, f) \log \frac{p(y, f)}{p(y)\, p(f)}\, df$

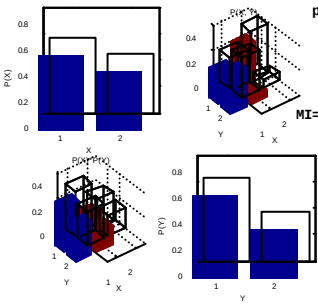
Mutual Information



$H(X,Y) = H(X) + H(Y) - MI(X,Y)$

$MI(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

MI Estimation

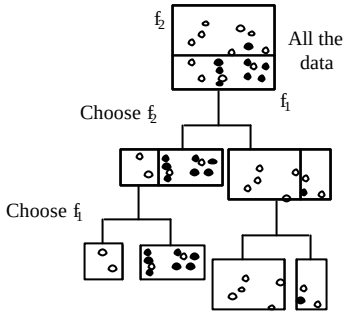


```
pxy=[4/16, 6/16; 5/16, 1/16]
px=sum(pxy)
py=sum(pxy, 2)
pxpy=(py*px)
MI=sum(sum(pxy.*log2(pxy./pxpy)))

pxy =   0.2500   0.3750
        0.3125   0.0625

px =    0.5625   0.4375

py =    0.6250
        0.3750

pxpy =  0.3516   0.2734
        0.2109   0.1641

MI =    0.1381
```

Possible normalization by (H(X)+H(Y))/2

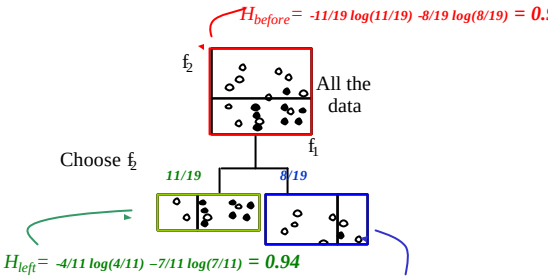
Tree Classifiers

CART (Breiman, 1984) or C4.5 (Quinlan, 1993)



At each step, choose the feature that "reduces entropy" most. Work towards "node purity".

Information Gain



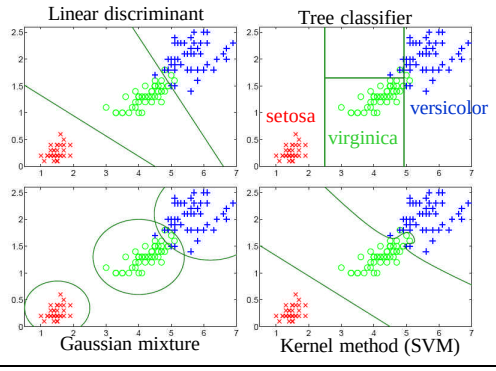
$IG = H_{before} - (11/19\ H_{left} + 8/19\ H_{right}) = 0.98 - 0.78 = 0.2$

IG = MI

- $H_{before} = H(Y)$
- $H_{left} = H(Y|f_2=-1)$
- $H_{right} = H(Y|f_2=1)$

- $IG = H_{before} - (p(f_2=-1)\ H_{left} + p(f_2=1)\ H_{right})$
  $= H(Y) - (p(f_2=-1)\ H(Y|f_2=-1) + p(f_2=1)\ H(Y|f_2=1))$
  $= H(Y) - H(Y|\ f_2)$
  $= MI(Y,f_2)$

Iris Data (Fisher, 1936/Chapter 1)



MI for Feature Selection

- Random Forest *(Breiman, Cutler, 2002)*
- Feature ranking
- Forward selection
- Feature subset selection (Markov Blankets, scaling factors, shrinkage)

Forward Selection with MI

Fleuret, 2004. *Practical only for binary features.*

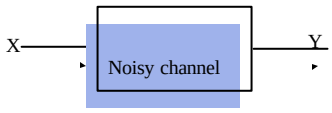
- Select a first feature $X_{?(1)}$ with maximum MI with the target.
- For each remaining feature X_i and each previously selected feature $X_{?(j)}$, compute the conditional mutual information:

$$CMI(X_i, Y \mid X_{?(j)}) = \Sigma_{X?(j)} P(X_{?(j)}) MI(X_i, Y \mid X_{?(j)})$$

- Select the feature with maximum CMI.

Transmission Channels

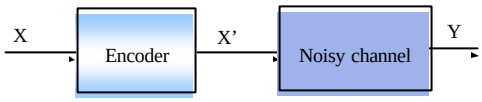
Transmission Channels



Channel capacity = $\max_{p(x)} MI(X, Y)$

= (second Shannon theorem) highest achievable information transmission rate without error.

Coding



Code efficiency = length(code) / length(original signal)

Theorems (Shannon, 1948):

- Errorless transmission is achievable with sufficiently inefficient (well chosen) codes; the efficiency is bounded by the capacity.
- For a chosen error rate, one can find a shorter code.

MI and Error Rate

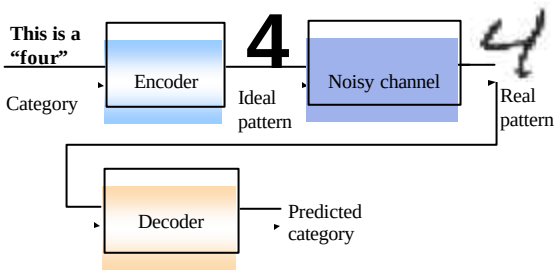
$$e_{bayes}(X) \geq 1 - \frac{I(Y,X) + \log 2}{\log(|Y|)}$$

Fano, 1961

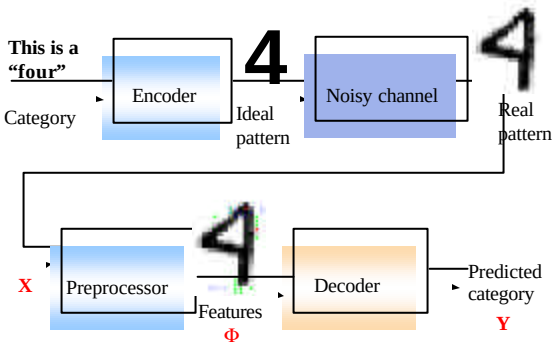
$$e_{bayes}(X) \leq \frac{1}{2} H(Y|X) = \frac{1}{2}(H(Y) - I(Y,X))$$

Hellman and Raviv (1970) Feder and Merhav (1994)

Noisy Channel in PR



Noisy Channel in PR



Information Bottleneck

Tishby et al., 1999

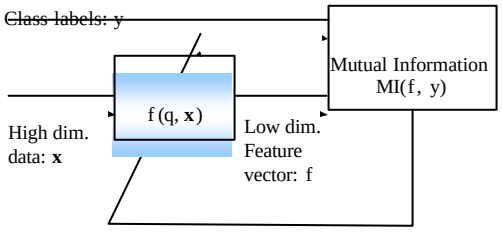
maximize MI(X,Y | Φ) = MI(Φ,Y)

but

minimize MI(X, Φ)

Information Bottleneck Methods

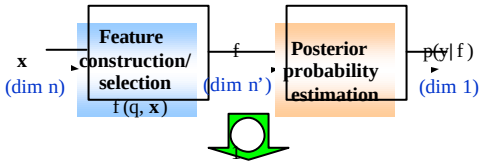
Gradient Descent



Gradient: ∂MI/∂q

Torkkola, 2001

Computational Tricks (1/4)



$$\frac{\partial MI}{\partial q}_{(1,\,n_\theta)} = \frac{\partial MI}{\partial f}_{(1,\,n')} \frac{\partial f}{\partial q}_{(n',\,n_\theta)}$$

p(y| f) appears here

For example $f_{(1,\,n')}(\mathbf{W}, \mathbf{x}) = \mathbf{W}_{(n',\,n)}\,\mathbf{x}^{T}_{(n,\,1)}$ (linear model) $\Rightarrow \partial f / \partial \mathbf{w} = \mathbf{x}$

Computational Tricks (2/4)

$$MI(Y, \Phi) = \sum_y \int_\phi p(y, f) \log \frac{p(y, f)}{p(y)\, p(f)}\, df$$

2

$$MI(Y, \Phi) = \sum_y \int_\phi (p(y, f) - p(y)\, p(f))^2\, df$$

Justification: Renyi entropy $H(Y) = -\log \sum_y p(y)^2$

Computational Tricks (3/4)

$$G(\mathbf{y}, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}^T \Sigma^{-1} \mathbf{y}\right)$$

$$\int_{\mathbf{y}} G(\mathbf{y} - \mathbf{a}_i, \Sigma_1) G(\mathbf{y} - \mathbf{a}_j, \Sigma_2) d\mathbf{y} = G(\mathbf{a}_i - \mathbf{a}_j, \Sigma_1 + \Sigma_2)$$

3

Use densities that are mixtures of Gaussians.

Examples of Suitable Densities

Gaussian mixture:

$p(f|y) = G(f - f_y, \Sigma_y)$

$p(f) = \sum_y p(y)\, p(f|y) = \sum_y (m_y/m)\, G(f - f_y, \Sigma_y)$

one per class

Parzen windows:

$p(f|y) = (1/m_y) \sum_{c \in y} G(f - f_c, \sigma)$

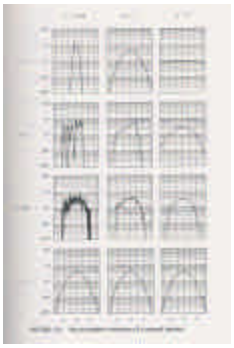
$p(f|y) = (1/m) \sum_i G(f - f_i, \sigma)$

one per example — the same for all

Parzen Windows (Rosenblatt 1956, Parzen 1962)

$p(x) = (1/m) \sum_{i=1:m} G(f - f_i, \sigma_m)$

$\sigma_m = \sigma/\mathrm{sqrt}(m)$



From Duda and Hart, 1974

Computational Tricks (4/4)

$$MI(Y, \Phi) = \sum_y \int_\phi (p(y, f) - p(y)\,p(f))^2 \, df$$

4

$$MI(Y, \Phi) = \sum_y \int_\phi p(y, f)^2 \, df \quad V_{in}$$

$$+ \sum_y \int_\phi p(y)^2\, p(f)^2 \, df \quad V_{all}$$

$$- 2 \sum_y \int_\phi p(y, f)\, p(y)\, p(f) \, df \quad V_{bw}$$

Plugging In

$$MI(Y, \Phi) = V_{in} + V_{all} - V_{bw}$$

V_{in} = function ($G(f_{ci} - f_{cj}, 2\sigma)$)

V_{all} = function ($G(f_i - f_j, 2\sigma)$)

V_{bw} = function ($G(f_{ci} - f_j, 2\sigma)$)

Information Forces

- Max MI by hill climbing (gradient descent):

$$q(t+1) = q(t) + \eta\, \partial MI/\partial q$$

$$= q(t) + \eta\, \partial MI/\partial f \;\; \partial f/\partial q$$

$$= q(t) + \eta\, [\partial V_{in}/\partial f + \partial V_{all}/\partial f - \partial V_{bw}/\partial f]\; \partial f/\partial q$$

- Force = - gradient (Energy) Energy $\Leftrightarrow$ - MI
- $\partial V_{in}/\partial f$ **Cohesion within classes**
- $\partial V_{all}/\partial f$ **Overall cohesion**
- $-\partial V_{bw}/\partial f$ **Repulsion to other classes**

Feature Construction

http://members.cox.net/torkkola/mmi, Torkkola, 2001

Example 1: Three classes in 3-dimensional space. Artificial data.

1.a f = linear, p(y|f) = Parzen windows

1.b f = RBF, p(y|f) = Parzen windows

Example 2: Three classes in 12-dimensional space. This is the oil-pipeflow data from NCRG at Aston University.

2.a f = linear, p(y|f) = Parzen windows

2.b f = linear, p(y|f) = Gaussian mixture

Example 3: Six classes in 36-dimensional space. This is the Landsat satellite image database from Statlog-project. f = linear, p(y|f) = Parzen windows

Application to Feature Selection

$\Phi_i = \sigma_i x_i$ scaling factors, $\sigma_i \geq 0$

$s^* = \text{argmax}_s \; MI(\Phi, Y)$

subject to: $\Sigma_i \; \sigma_i^2 = 1$ (shear)

$s^* = \text{argmin}_s \; MI(X, \Phi) - \beta \, MI(\Phi, Y)$

Conclusion

- MI characterizes the **dependency between random** variables:
  - Feature ranking
  - Forward selection with CMI
- MI is also an **Information Gain**
  - Tree classifiers
  - Random Forests
- MI characterizes the **channel capacity** and the **rate of distorsion**
  - Information bottleneck methods

    $\min_q \; MI(X, \Phi) - \beta \; MI(\Phi, Y)$

Exercise Class

Madelon

How to review a paper?

Exercises on IT methods

Madelon (continued)

```
my_classif=svc({'coef0=1', 'degree=0', 'gamma=1',
'shrinkage=1'});
my_model=chain({probe(relief,{'p_num=2000',
'pval_max=0'}), standardize, my_classif})
'f_max=20'
```

Earn 1 point with a result better than the baseline method (testBER<7.33 or [testBER=7.33 and feat_num<20 (4%)]).

Earn a total of 2 points with a result testBER<6.79% or [testBER=6.79% and feat_num<17 (3.4%)].

Homework

- Write a review of a paper in the "Feature Extraction" book (the paper you presented or will present or any paper/chapter you want).
- Questions:
  - What is the goal of a review?
  - What is the goal of publishing?

Review form

A. Eliminating criteria (a zero means rejection)
1. Scope
2. Novelty
3. Usefulness
4. Sanity

B. Other fundamental criteria (a low grade implies major revisions)
1. Quantity
2. Reproducibility
3. Demonstration
4. Comparison

Review Form (continued)

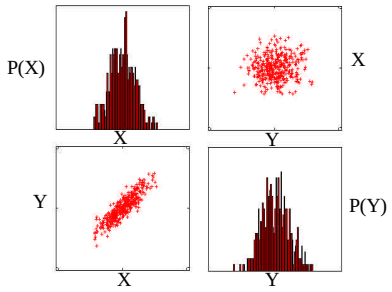
C. Forgivable problems of contents (minor revisions may be required)
1. Completeness
2. Take-aways
3. Bibliography
4. Outlook

D. Bonus reproducibility criteria (give additional credit)
1. Data availability
2. Code availability

Review Form (end)

E. Message delivery (major/minor revisions may be required)
1. Readability
2. Notations
3. Figures
4. Formalism
5. Density
6. Language

MI and Pearson Correlation



MI(X, Y) = -(1/2) log(1-R^2)

Exercise: prove MI=-1/2 log(1-R^2)

Prove:

- Entropy Gaussian source: $H(X) = \frac{1}{2} \log(2\pi e)\sigma^2$
- Solution of one-variable least-square regression y=ax: $a=\sigma_{xy}/\sigma_x^2$. We will call a' the sol. of x=a'y.
- Pearson correlation coefficient: $R^2=aa'$
- Residual error $\sigma^2 = E(y-ax)^2 = (1-R^2)\sigma_y^2$.
- Y=function(X)+N, N is random noise
  - MI(X,Y)=H(Y)-H(N)
  - MI(X,Y)≥H(X)-H(N)
- For a Gaussian channel with Gaussian source: $MI(X,Y)=-\frac{1}{2} \log \sigma^2/\sigma_y^2$. Conclude the proof.

Markov Blankets

- Notations: All features X, Features selected Φ, complement Φ', X=[Φ, Φ'].
- Definition: A **Markov Blanket** is a subset of features Φ such that:

$$P(X,Y|\Phi) = P(X|\Phi)\ P(Y|\Phi)$$

We write $X \perp Y \mid \Phi$ (X is independent of Y given Φ)

Exercise (Markov Blankets)

- If Φ is a MB, prove $P(Y|\Phi) = P(Y|X)$ for all assigments of values to Φ'.
- We define $DMI(\Phi) = E_{p(X)}\ D_{KL}(P(Y|X) \mid\mid P(Y|\Phi))$
  Prove: $DMI(\Phi)=E_{p(X,Y)}\ P(Y|X)/P(Y|\Phi)$
- We define $A=\log[P(X,Y)/(P(X)P(Y))]$ and $B=\log[P(\Phi,Y)/(P(\Phi)P(Y))]$, prove:
  $A-B=P(Y|X)/P(Y|\Phi)$
  $\Sigma_{\{\Phi Y\}}P(\Phi,Y)B= \Sigma_{\{XY\}}P(X,Y)B$
  $DMI(\Phi) = MI(X,Y) - MI(\Phi,Y)$
- Define an optimization problem to find a minimum MB. How can you relate it to the information bottleneck approach?

Complement of exercise

- What are the similarities and differences between the “bottleneck” IT methods and the Bayesian/regularized risk methods?
- Can you suggest hybrid approaches?
- Can you suggest forward selection and backward elimination IT methods using the scaling factor approach?