Information theoretic feature selection in multi-label data through composite likelihood

Konstantinos Sechidis, Nikolaos Nikolaou, and Gavin Brown

School of Computer Science
University of Manchester

The University of Manchester

Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to > 1 labels

Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to > 1 labels
- Applications

Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to > 1 labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions

Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to > 1 labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions
  - Text Mining: 1 webpage/document, many categories

Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to > 1 labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions
  - Text Mining: 1 webpage/document, many categories
  - Image Retrieval: 1 image, many semantic concepts

Male, Person,
Motorbike, Vehicle
Building

Female, Person,
Building

Male, Person

Rabbit, Animal
Car, Vehicle

Motivation: Multi-label Learning

- Multi-label: Each datapoint can be associated to > 1 labels
- Applications
  - Bioinformatics: 1 gene/protein, many functions
  - Text Mining: 1 webpage/document, many categories
  - Image Retrieval: 1 image, many semantic concepts

Male, Person, Motorbike, Vehicle Building

Female, Person, Building

Male, Person

Rabbit, Animal Car, Vehicle

- Common characteristic of these domains: Large number of features

Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information

Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information
- Filters: Functions that assign a "utility" score to each feature

Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information
- Filters: Functions that assign a “utility” score to each feature
- In this work we discuss information-theoretic filters

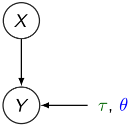
Feature Selection

- Feature Selection: Find minimal subset of features with maximal useful information
- Filters: Functions that assign a “utility” score to each feature
- In this work we discuss information-theoretic filters
- Filter Assumption: model and feature selection are independent

Feature Selection via Likelihood Maximization

- Brown et al. (JMLR 2012) unified many heuristic information-theoretic filter criteria for feature selection

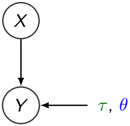
Conditional Likelihood Maximization under model



Feature Selection via Likelihood Maximization

- Brown et al. (JMLR 2012) unified many heuristic information-theoretic filter criteria for feature selection

  Conditional Likelihood Maximization under model



- Negative log-likelihood asymptotically decomposes into 3 terms:

$$\lim_{N\to\infty} -\ell = \text{model term} + \text{feature selection term} + \text{Bayes error}$$

  feature selection is mutual info $I(X_\theta; Y)$

Single-label Feature Selection Criteria

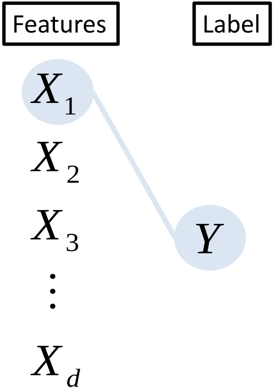
Feature space independence assumptions:

Full:

Features	Label
X_1	
X_2	
X_3	Y
$\vdots$	
X_d	

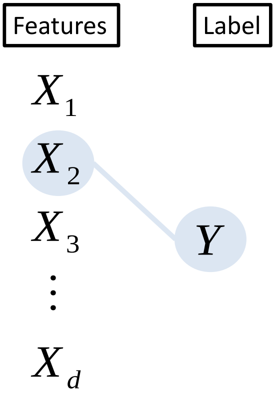
Single-label Feature Selection Criteria

Feature space independence assumptions:
Full:



Feature space independence assumptions:

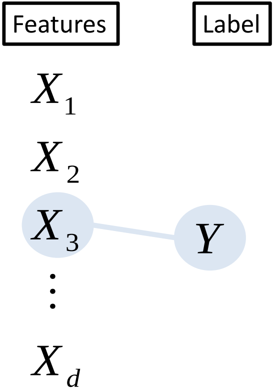
Full:



Single-label Feature Selection Criteria

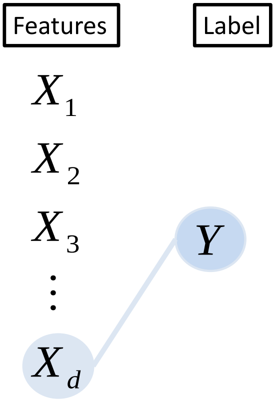
Feature space independence assumptions:

Full:



Feature space independence assumptions:

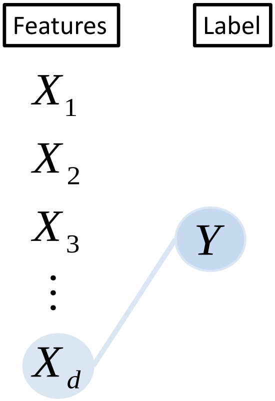
Full:



Single-label Feature Selection Criteria

Feature space independence assumptions:

Full:



$J_{MIM}(X_k) = I(X_k; Y)$

Single-label Feature Selection Criteria

Feature space independence assumptions:

Full:

Features

Label

X_1

X_2

X_3

$\vdots$

X_d

Y

$$J_{MIM}(X_k) = I(X_k; Y)$$

Partial
(i.e. pairwise dependencies):

Features

Label

X_1

X_2

X_3

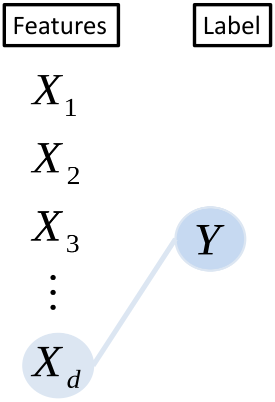
$\vdots$

X_d

Y

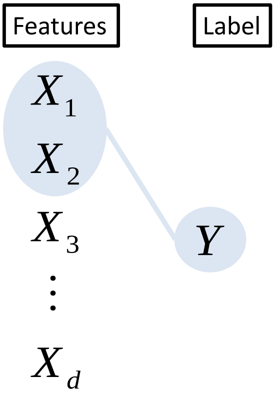
Feature space independence assumptions:

Full:



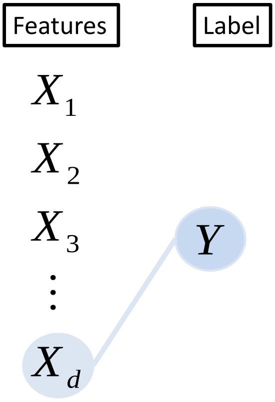
$$J_{MIM}(X_k) = I(X_k; Y)$$

Partial
(i.e. pairwise dependencies):



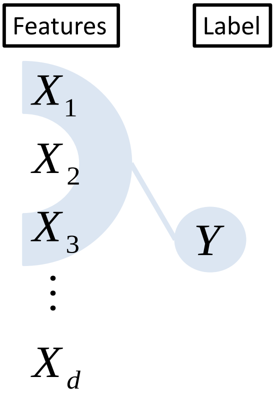
Feature space independence assumptions:

Full:



$$J_{MIM}(X_k) = I(X_k; Y)$$

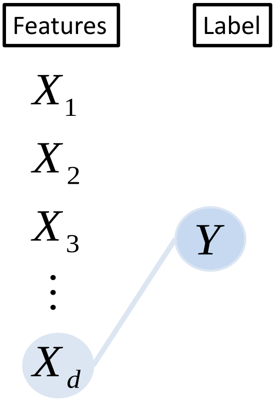
Partial
(i.e. pairwise dependencies):



Single-label Feature Selection Criteria

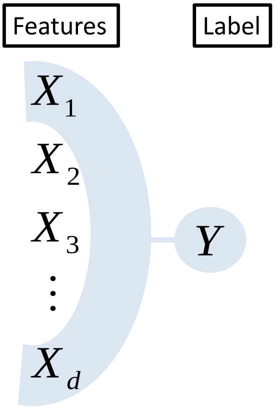
Feature space independence assumptions:

Full:



$$J_{MIM}(X_k) = I(X_k; Y)$$

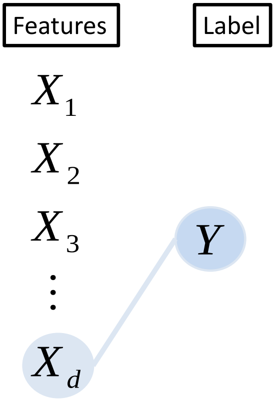
Partial
(i.e. pairwise dependencies):



Single-label Feature Selection Criteria

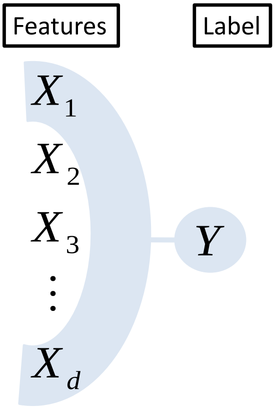
Feature space independence assumptions:

Full:



$$J_{MIM}(X_k) = I(X_k; Y)$$

Partial
(i.e. pairwise dependencies):

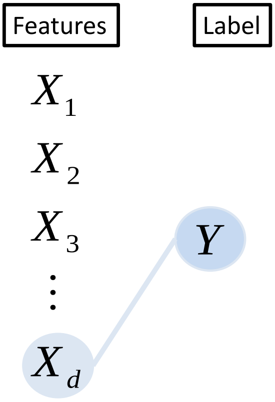


$$J_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_{\theta_j} X_k; Y)$$

Single-label Feature Selection Criteria

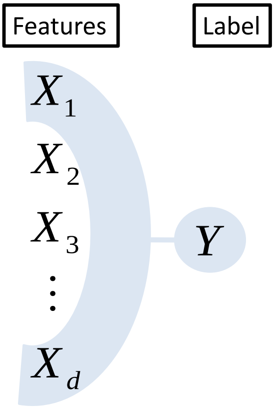
Feature space independence assumptions:

Full:



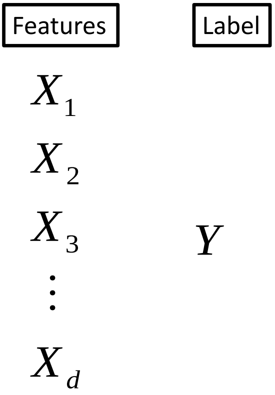
$$J_{MIM}(X_k) = I(X_k; Y)$$

Partial
(i.e. pairwise dependencies):



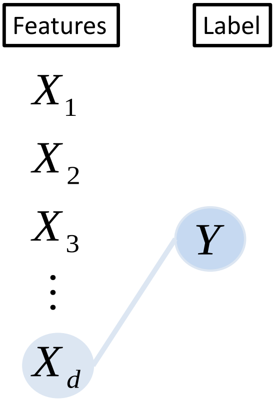
$$J_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_{\theta_j} X_k; Y)$$

None:



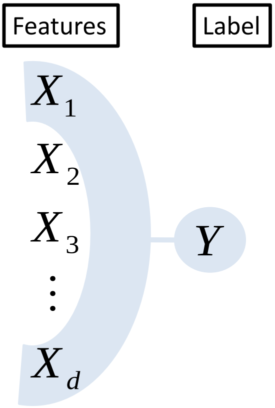
Feature space independence assumptions:

Full:



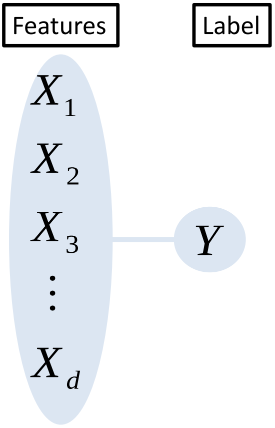
$$J_{MIM}(X_k) = I(X_k; Y)$$

Partial
(i.e. pairwise dependencies):



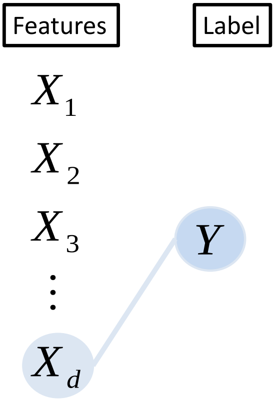
$$J_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_{\theta_j} X_k; Y)$$

None:



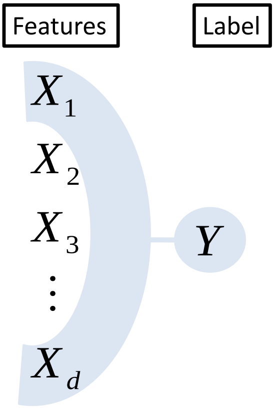
Feature space independence assumptions:

Full:



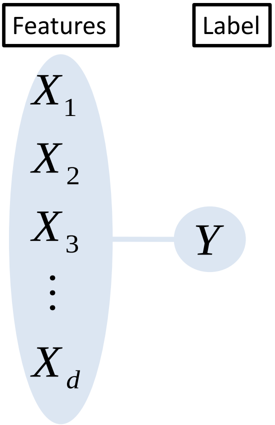
$$J_{MIM}(X_k) = I(X_k; Y)$$

Partial
(i.e. pairwise dependencies):



$$J_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_{\theta_j} X_k; Y)$$

None:



$$J_{CMI}(X_k) = I(X_k; Y | X_\theta)$$

Extending Framework to Multi-label Setting

- Next, extend to multi-label where Y is q-dimensional

Man, Hat,
Person

Extending Framework to Multi-label Setting

- Next, extend to multi-label where Y is q-dimensional

Man, Hat,
Person

- What independence assumptions can we make in label space?

Extending Framework to Multi-label Setting

- Next, extend to multi-label where Y is q-dimensional

Man, Hat,
Person

- What independence assumptions can we make in label space?
- In this work we examined:
  - Binary Relevance (BR) vs Label Powerset (LP)

Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels

Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary q-label problem $\Rightarrow$ 1 single-label, 2^q-class problem

Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary q-label problem $\Rightarrow$ 1 single-label, 2^q-class problem

	Animal	Building	Vehicle		y
	1	0	1	$\Rightarrow$	1 0 1
	0	1	1	$\Rightarrow$	0 1 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary q-label problem $\Rightarrow$ 1 single-label, 2^q-class problem

	Animal	Building	Vehicle		y
	1	0	1	$\Rightarrow$	1 0 1
	0	1	1	$\Rightarrow$	0 1 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

- Pros: dependencies among labels are accounted for

Multi-label Extension: LP Transformation

- Label Powerset (LP): No independence among labels
- Binary q-label problem $\Rightarrow$ 1 single-label, 2^q-class problem

	Animal	Building	Vehicle		y
	1	0	1	$\Rightarrow$	1 0 1
	0	1	1	$\Rightarrow$	0 1 1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

- Pros: dependencies among labels are accounted for
- Cons: probability estimates unreliable (curse of dimensionality)

Multi-label Extension: LP Transformation

Feature space independence assumptions:

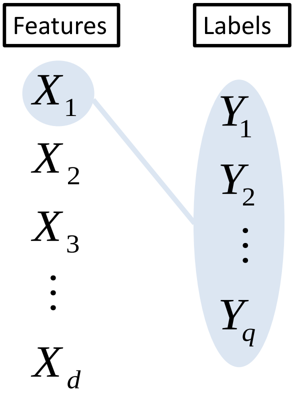
Full:

Features	Labels
X_1	Y_1
X_2	Y_2
X_3	$\vdots$
$\vdots$	Y_q
X_d	

Multi-label Extension: LP Transformation

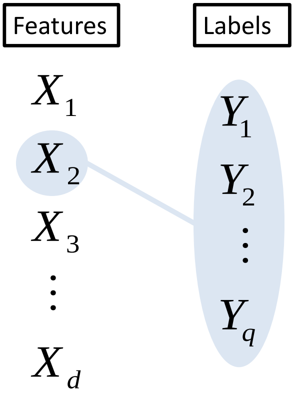
Feature space independence assumptions:

Full:



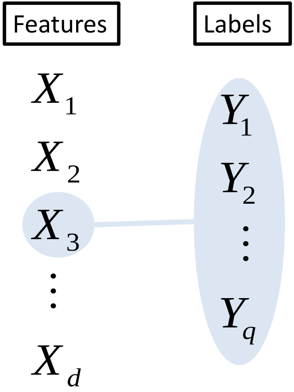
Feature space independence assumptions:

Full:



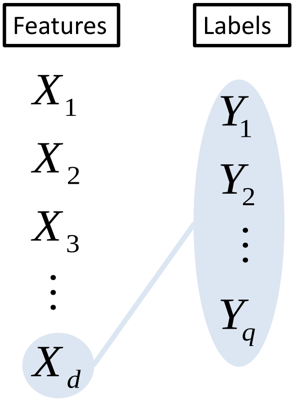
Feature space independence assumptions:

Full:



Feature space independence assumptions:

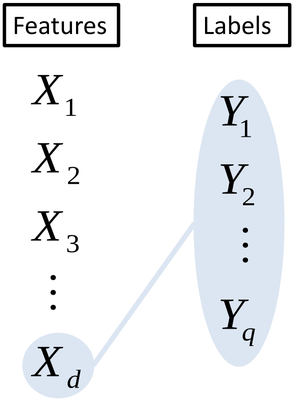
Full:



Multi-label Extension: LP Transformation

Feature space independence assumptions:

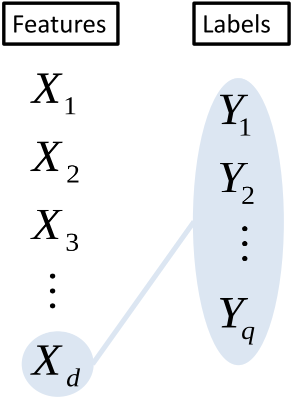
Full:



$$J^{LP}_{MIM}(X_k) = I(X_k; Y_{1:q})$$

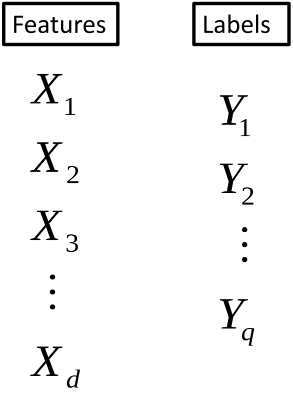
Feature space independence assumptions:

Full:



$$J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})$$

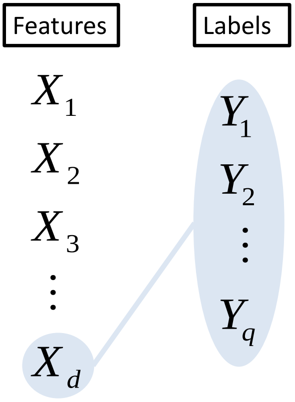
Partial
(i.e. pairwise dependencies):



Multi-label Extension: LP Transformation

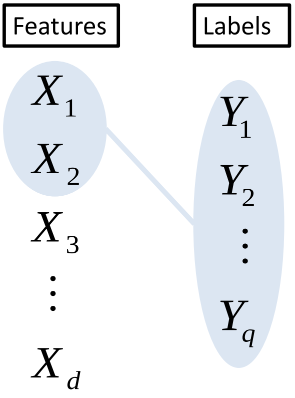
Feature space independence assumptions:

Full:



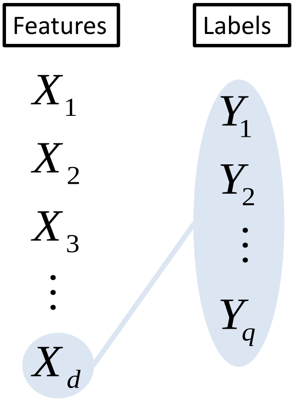
$J^{LP}_{MIM}(X_k) = I(X_k; Y_{1:q})$

Partial
(i.e. pairwise dependencies):



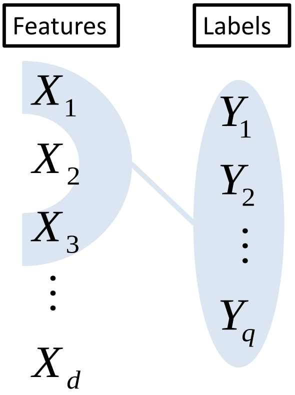
Feature space independence assumptions:

Full:



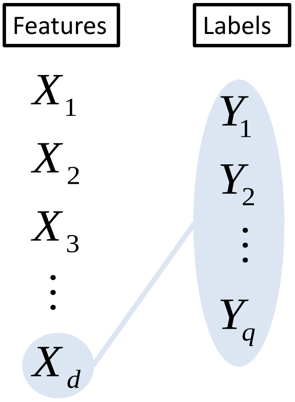
$J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})$

Partial
(i.e. pairwise dependencies):



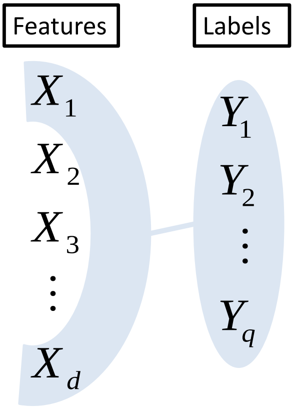
Feature space independence assumptions:

Full:



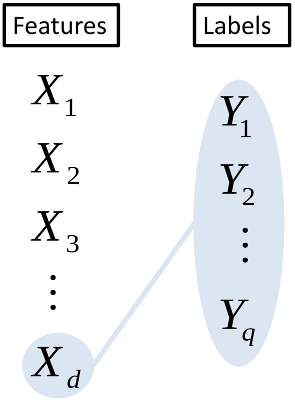
$$J^{LP}_{MIM}(X_k) = I(X_k; Y_{1:q})$$

Partial
(i.e. pairwise dependencies):



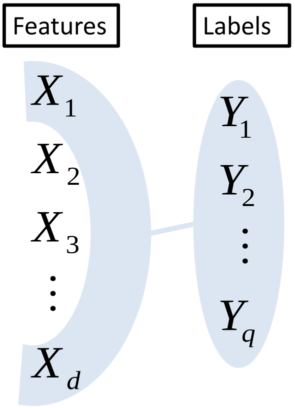
Feature space independence assumptions:

Full:



$$J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})$$

Partial
(i.e. pairwise dependencies):

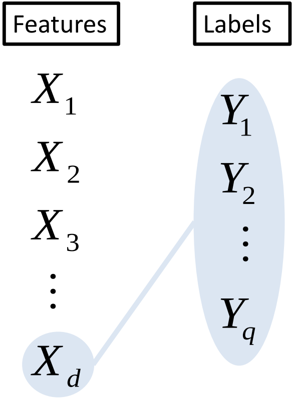


$$J_{JMI}^{LP}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_k X_{\theta_j}; Y_{1:q})$$

Multi-label Extension: LP Transformation

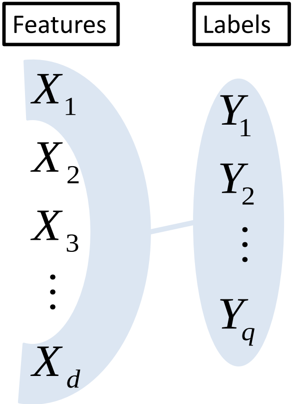
Feature space independence assumptions:

Full:



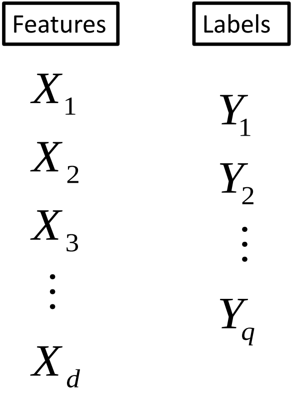
$$J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})$$

Partial
(i.e. pairwise dependencies):



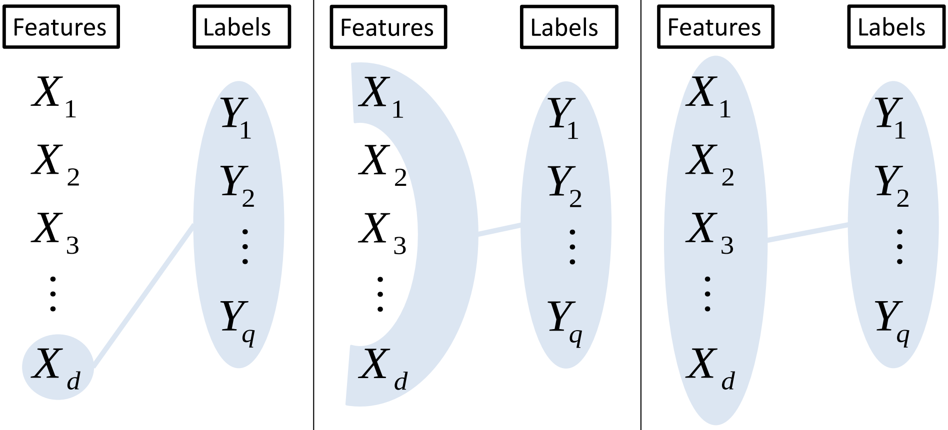
$$J_{JMI}^{LP}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_k X_{\theta_j}; Y_{1:q})$$

None:



Feature space independence assumptions:

Full: | Partial (i.e. pairwise dependencies): | None:

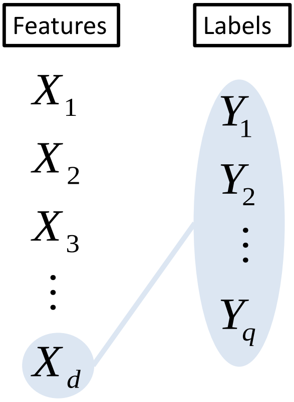


$$J_{MIM}^{LP}(X_k) = I(X_k; Y_{1:q})$$

$$J_{JMI}^{LP}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_k X_{\theta_j}; Y_{1:q})$$

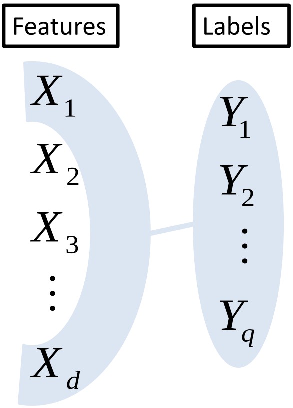
Feature space independence assumptions:

Full:



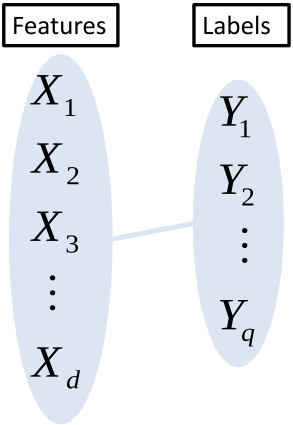
$$J^{LP}_{MIM}(X_k) = I(X_k; Y_{1:q})$$

Partial
(i.e. pairwise dependencies):



$$J^{LP}_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} I(X_k X_{\theta_j}; Y_{1:q})$$

None:



$$J^{LP}_{CMI}(X_k) = I(X_k; Y_{1:q} | X_\theta)$$

Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels

Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary q-label problem $\Rightarrow$ q independent single-label, binary problems

Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary q-label problem $\Rightarrow$ q independent single-label, binary problems

	Animal	Building	Vehicle		y_1	y_2	y_3
	1	0	1	$\Rightarrow$	1	0	1
	0	1	1	$\Rightarrow$	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$

Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary q-label problem $\Rightarrow$ q independent single-label, binary problems

	Animal	Building	Vehicle		y_1	y_2	y_3
	1	0	1	$\Rightarrow$	1	0	1
	0	1	1	$\Rightarrow$	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$

- Pros: more reliable probability estimates

Multi-label Extension: BR Transformation

- Binary Relevance (BR): Full Independence among labels
- Binary q-label problem $\Rightarrow$ q independent single-label, binary problems

	Animal	Building	Vehicle		y_1	y_2	y_3
	1	0	1	$\Rightarrow$	1	0	1
	0	1	1	$\Rightarrow$	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$

- Pros: more reliable probability estimates
- Cons: dependencies among labels are not accounted for

Multi-label Extension: BR Transformation

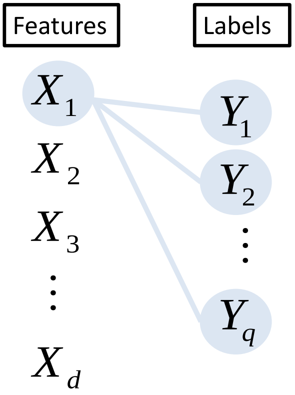
Feature space independence assumptions:

Full:

Features	Labels
X_1	Y_1
X_2	Y_2
X_3	$\vdots$
$\vdots$	Y_q
X_d	

Feature space independence assumptions:

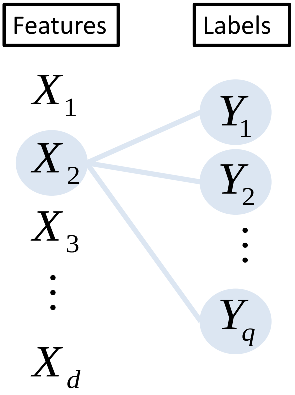
Full:



Multi-label Extension: BR Transformation

Feature space independence assumptions:

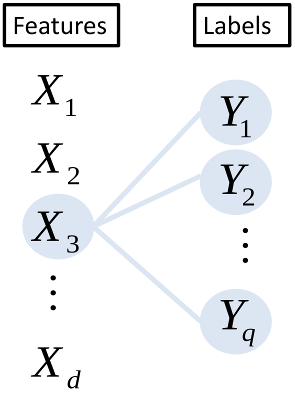
Full:



Multi-label Extension: BR Transformation

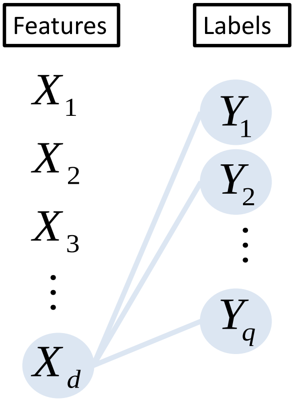
Feature space independence assumptions:

Full:

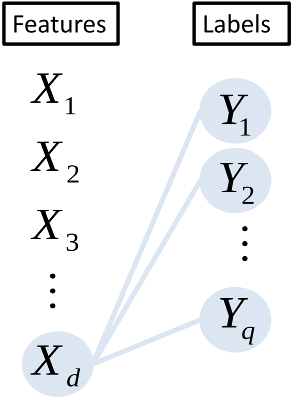


Feature space independence assumptions:

Full:



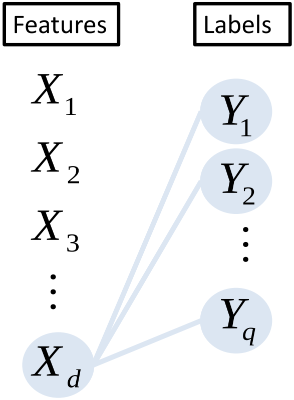
Feature space independence assumptions:
Full:



$$J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

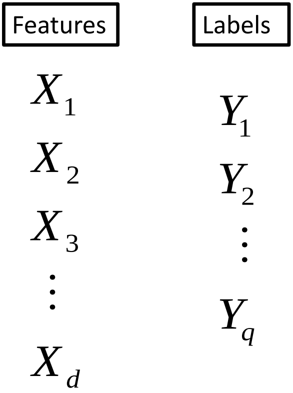
Feature space independence assumptions:

Full:



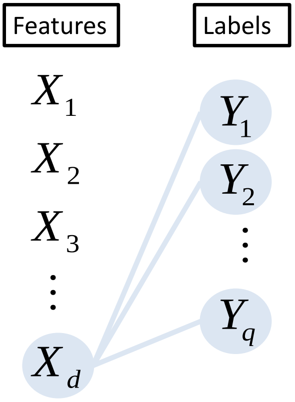
$$J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

Partial
(i.e. pairwise dependencies):



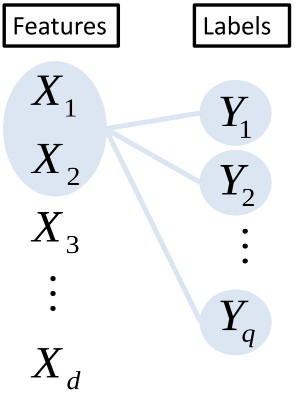
Feature space independence assumptions:

Full:



$$J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

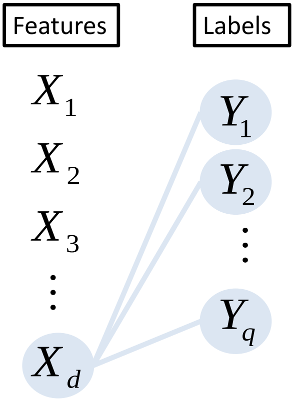
Partial
(i.e. pairwise dependencies):



Multi-label Extension: BR Transformation

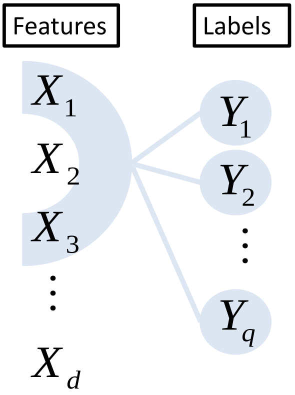
Feature space independence assumptions:

Full:



$$J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

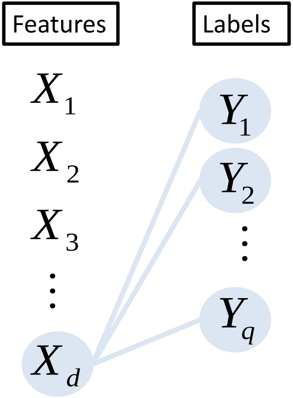
Partial
(i.e. pairwise dependencies):



Multi-label Extension: BR Transformation

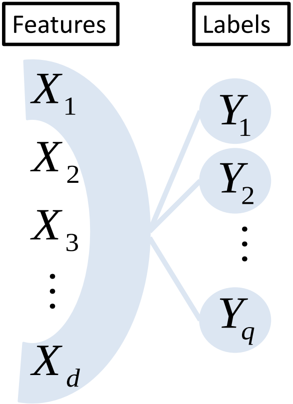
Feature space independence assumptions:

Full:



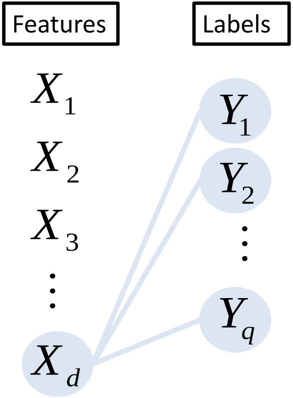
$$J_{MIM}^{BR}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

Partial
(i.e. pairwise dependencies):



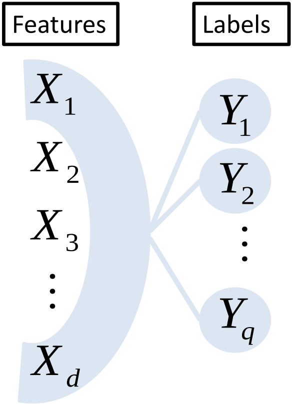
Feature space independence assumptions:

Full:



$$J^{BR}_{MIM}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

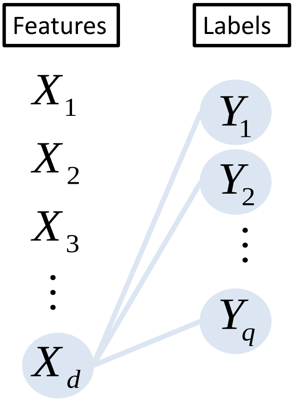
Partial (i.e. pairwise dependencies):



$$J^{BR}_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} \sum_{l=1}^{q} I(X_k X_{\theta_j}; Y_l)$$

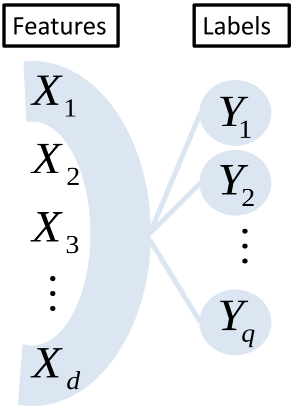
Feature space independence assumptions:

Full:



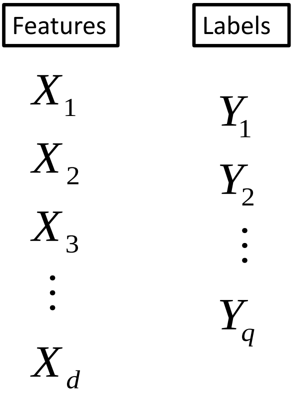
$$J^{BR}_{MIM}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

Partial
(i.e. pairwise dependencies):



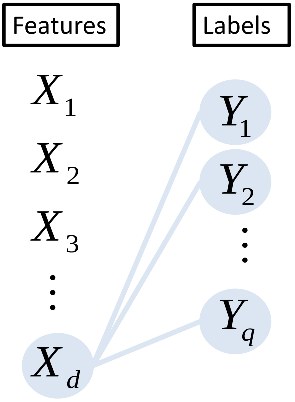
$$J^{BR}_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} \sum_{l=1}^{q} I(X_k X_{\theta_j}; Y_l)$$

None:



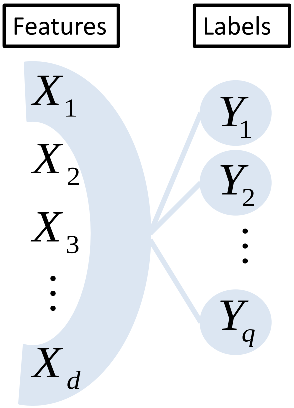
Feature space independence assumptions:

Full:



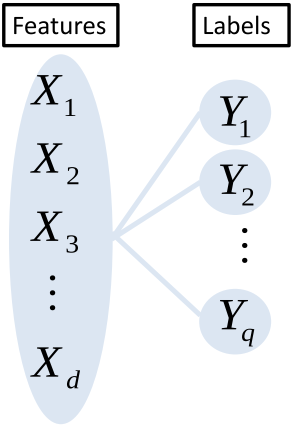
$$J^{BR}_{MIM}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

Partial
(i.e. pairwise dependencies):



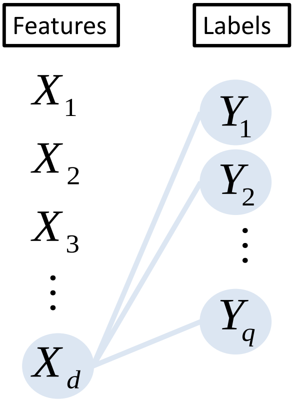
$$J^{BR}_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} \sum_{l=1}^{q} I(X_k X_{\theta_j}; Y_l)$$

None:



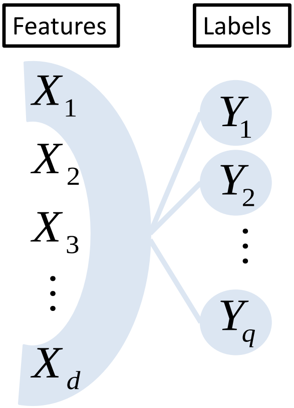
Feature space independence assumptions:

Full:



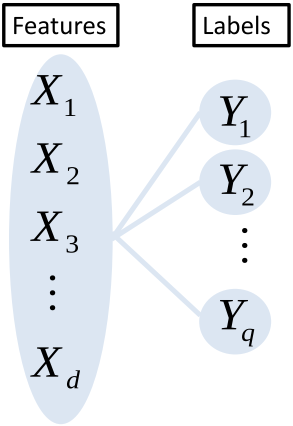
$$J^{BR}_{MIM}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l)$$

Partial
(i.e. pairwise dependencies):



$$J^{BR}_{JMI}(X_k) = \sum_{j=1}^{|X_\theta|} \sum_{l=1}^{q} I(X_k X_{\theta_j}; Y_l)$$

None:



$$J^{BR}_{CMI}(X_k) = \sum_{l=1}^{q} I(X_k; Y_l | X_\theta)$$

Existing and New Criteria under our Framework

- Summarizing, the criteria based on feature space X and label space Y independence assumptions:

Label space independence assumptions	Feature space independence assumptions: *CMI* (none)	*JMI* (partial)	*MIM* (full)
Label Powerset (none)	$J^{\mathrm{Y:none}}_{\mathrm{X:none}}$	$J^{\mathrm{Y:none}}_{\mathrm{X:partial}}$	$J^{\mathrm{Y:none}}_{\mathrm{X:full}}$
Binary Relevance (full)	$J^{\mathrm{Y:full}}_{\mathrm{X:none}}$	$J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$	$J^{\mathrm{Y:full}}_{\mathrm{X:full}}$

Existing and New Criteria under our Framework

- Summarizing, the criteria based on feature space X and label space Y independence assumptions:

Label space independence assumptions	Feature space independence assumptions: *CMI* (none)	*JMI* (partial)	*MIM* (full)
Label Powerset (none)	*Doquire & Verleysen (2013)*	$J^{\mathrm{Y:none}}_{\mathrm{X:partial}}$	$J^{\mathrm{Y:none}}_{\mathrm{X:full}}$
Binary Relevance (full)	$J^{\mathrm{Y:full}}_{\mathrm{X:none}}$	$J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$	$J^{\mathrm{Y:full}}_{\mathrm{X:full}}$

Existing and New Criteria under our Framework

- Summarizing, the criteria based on feature space X and label space Y independence assumptions:

Label space independence assumptions		Feature space independence assumptions		
		CMI (none)	*JMI* (partial)	*MIM* (full)
	Label Powerset (none)	*Doquire & Verleysen (2013)*	$J^{\text{Y:none}}_{\text{X:partial}}$	*Spolaôr et al. (2013)*
	Binary Relevance (full)	$J^{\text{Y:full}}_{\text{X:none}}$	$J^{\text{Y:full}}_{\text{X:partial}}$	$J^{\text{Y:full}}_{\text{X:full}}$

- Summarizing, the criteria based on feature space X and label space Y independence assumptions:

Label space independence assumptions	Feature space independence assumptions: *CMI* (none)	*JMI* (partial)	*MIM* (full)
Label Powerset (none)	*Doquire & Verleysen (2013)*	$J^{\mathrm{Y:none}}_{\mathrm{X:partial}}$	*Spolaôr et al. (2013)*
Binary Relevance (full)	$J^{\mathrm{Y:full}}_{\mathrm{X:none}}$	$J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$	*Young & Pedersen (1997), Trohidis et al. (2008), ...*

Experiments

- Compare

Experiments

- Compare
  - effect of label space assumptions

Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions

Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art

Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art
- Procedure: Select M top features under each criterion, classify, evaluate; vary M

Experiments

- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art
- Procedure: Select M top features under each criterion, classify, evaluate; vary M
- Datasets: scene and yeast

Experiments

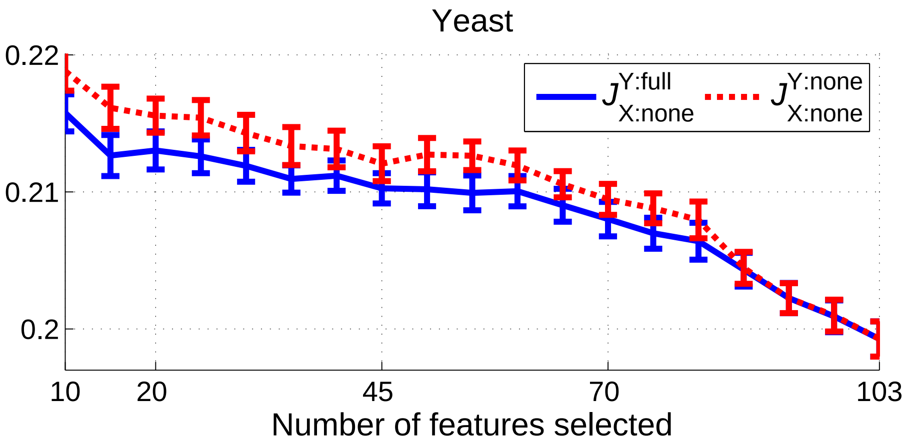
- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art
- Procedure: Select M top features under each criterion, classify, evaluate; vary M
- Datasets: scene and yeast
- Classification: ML-kNN, $k = 7$

Experiments

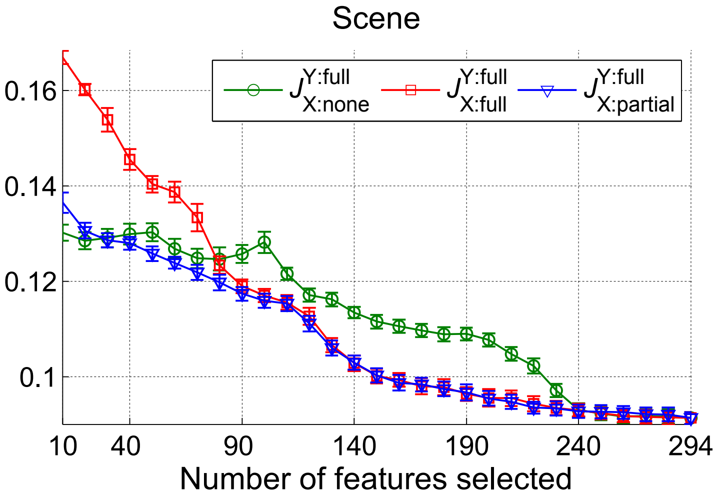
- Compare
  - effect of label space assumptions
  - effect of feature space assumptions
  - our best criterion vs. state-of-the-art
- Procedure: Select M top features under each criterion, classify, evaluate; vary M
- Datasets: scene and yeast
- Classification: ML-kNN, $k = 7$
- Evaluation: Hamming Loss (shown here) and Ranking Loss (similar)

Effect of Label Space Assumptions

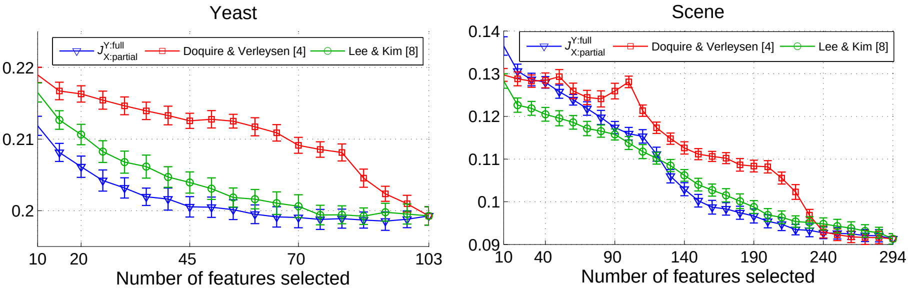
Hamming loss



Hamming loss

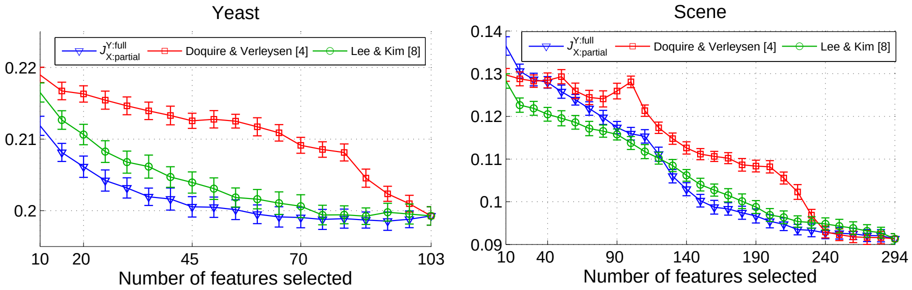


Hamming loss



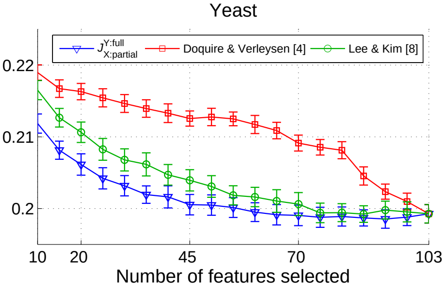
$J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$ vs. State-of-the-art (1)

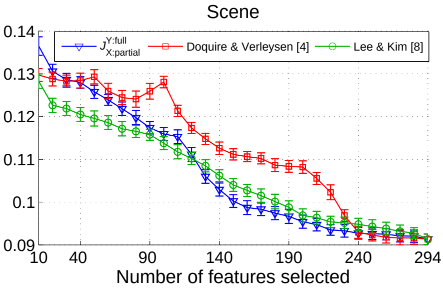
Hamming loss



- Doquire & Verleysen (2013) : $J^{\mathrm{Y:none}}_{\mathrm{X:none}}$ with pruning of rare cases

Hamming loss

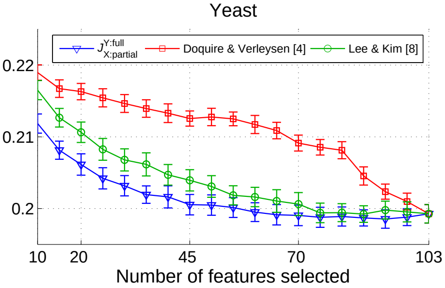


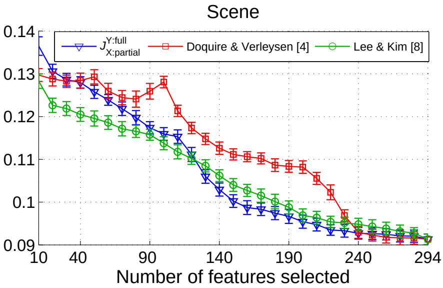


- Doquire & Verleysen (2013) : $J^{\mathrm{Y:none}}_{\mathrm{X:none}}$ with pruning of rare cases
- Lee & Kim (2013) : Multivariate Mutual Information

$J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$ vs. State-of-the-art (1)

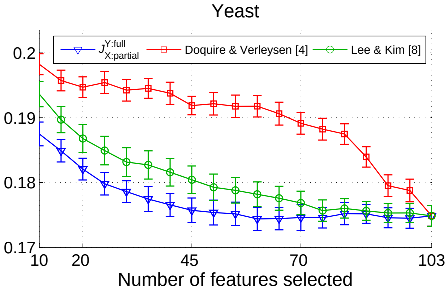
Hamming loss

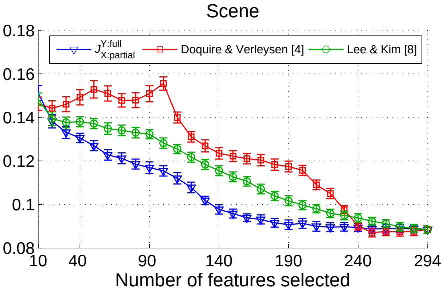




- Doquire & Verleysen (2013) : $J^{\mathrm{Y:none}}_{\mathrm{X:none}}$ with pruning of rare cases
- Lee & Kim (2013) : Multivariate Mutual Information
- $J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$ tends to outperform state-of-the-art criteria

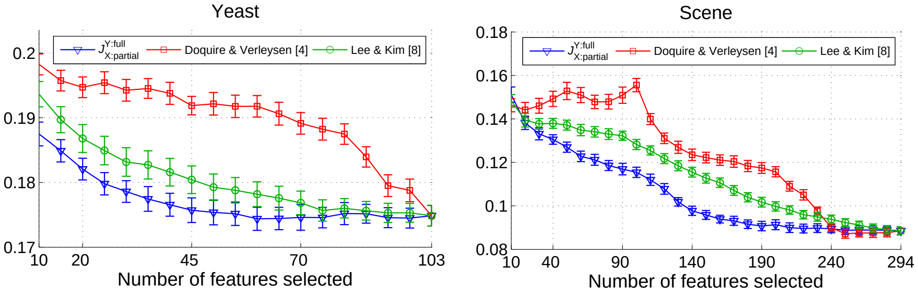
Ranking loss





- Doquire & Verleysen (2013) : $J^{\text{Y:none}}_{\text{X:none}}$ with pruning of rare cases
- Lee & Kim (2013) : Multivariate Mutual Information

Ranking loss



- Doquire & Verleysen (2013) : $J^{\mathrm{Y:none}}_{\mathrm{X:none}}$ with pruning of rare cases
- Lee & Kim (2013) : Multivariate Mutual Information
- $J^{\mathrm{Y:full}}_{\mathrm{X:partial}}$ dominates state-of-the-art criteria

Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...

Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space

Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers

Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
- ...in feature space, JMI gives best results

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
- ...in feature space, JMI gives best results
  - Examining pairwise interactions seems a good compromise between capturing interdependencies vs obtaining reliable estimates...

Empirical Observations

- Caution: Only 2 datasets! But based on them it appears that...
- ...independence assumptions in label space matter less than in feature space
  - Agrees with Gharroudi et al. (2014) for multilabel-label wrappers
- ...in feature space, JMI gives best results
  - Examining pairwise interactions seems a good compromise between capturing interdependencies vs obtaining reliable estimates...
  - Agrees with Brown et al. (JMLR 2012) findings in single-label filters

Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...

Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...
- ...as informative priors $P(X)$ or $P(Y)$

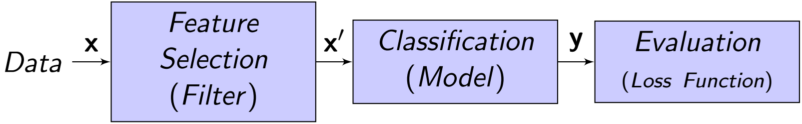
Future Work: Incorporating Domain Knowledge

- Probabilistic framework allows explicit incorporation of domain knowledge...
- ...as informative priors $P(X)$ or $P(Y)$
- ...as to how the distribution $P(Y|X)$ is factored

Future Work: Incorporating Domain Knowledge

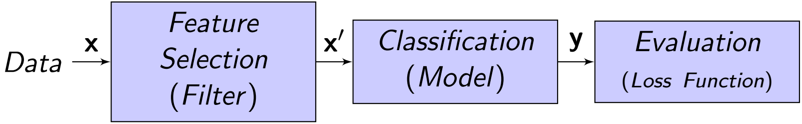
- Probabilistic framework allows explicit incorporation of domain knowledge...
- ...as informative priors $P(X)$ or $P(Y)$
- ...as to how the distribution $P(Y|X)$ is factored
- Thus constructing more problem specific filters

- A typical machine learning pipeline



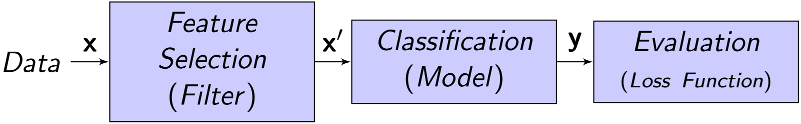
Future Work: The Bigger Picture

- A typical machine learning pipeline



- Assumptions in every step, often conflicting...

- A typical machine learning pipeline



- Assumptions in every step, often conflicting...
- ...should investigate interplay between model, filter & loss function

Thank you!
Kiitos!