Missing Data Problems in Machine Learning

Senate Thesis Defense

Ben Marlin
Machine Learning Group
Department of Computer Science
University of Toronto
April 8, 2008

Overview:

- Theory of Missing Data
- Unsupervised Learning - MAR
- Classification with Complete Data
- Unsupervised Learning - NMAR
- Classification with Missing Features
- Collaborative Prediction

Introduction: Notation for Missing Data

$\mathbf{x}_n$	0.1 | 0.9 | 0.2 | 0.7 | 0.3	Data Vector
$\mathbf{r}_n$	1 | 0 | 0 | 1 | 1	Response Vector
$\mathbf{o}_n$	1 | 4 | 5	Observed Dimensions
$\mathbf{m}_n$	2 | 3	Missing Dimensions
$\mathbf{x}_n^{\mathbf{o}_n}, \mathbf{x}_n^{o}$	0.1 | 0.7 | 0.3	Observed Data
$\mathbf{x}_n^{\mathbf{m}_n}, \mathbf{x}_n^{m}$	0.9 | 0.2	Missing Data

Theory of Missing Data: Factorizations

Data/Selection Model Factorization:

$$P(\mathbf{X},\mathbf{R},\mathbf{Z}|\theta,\mu) = P(\mathbf{R}|\mathbf{X},\mathbf{Z},\mu)P(\mathbf{X},\mathbf{Z}|\theta)$$

• The probability of selection depends on the true values of the data variables and latent variables.

Classification of Missing Data:

MCAR: $P(\mathbf{R}|\mathbf{X},\mathbf{Z},\mu) = P(\mathbf{R}|\mu)$

MAR: $P(\mathbf{R}|\mathbf{X},\mathbf{Z},\mu) \cancel{=} P(\mathbf{R}|X^o,\mu)$

NMAR: $P(\mathbf{R}|\mathbf{X},\mathbf{Z},\mu)$ No simplification in general.

Theory of Missing Data: Inference

MCAR/MAR Posterior:

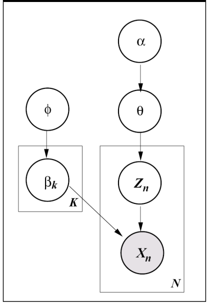
$$P(\theta|\mathbf{x}^o,\mathbf{r}) \propto \int\int\int P(\mathbf{X},\mathbf{Z}|\theta)P(\mathbf{R}|\mathbf{X},\mu)P(\theta|\omega)P(\mu|\eta)d\mu dZ d\mathbf{x}^m$$
$$\propto P(\mathbf{X}^o = \mathbf{x}^o|\theta)P(\theta|\omega)$$

NMAR Posterior:

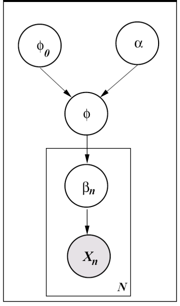
$$P(\theta|\mathbf{x}^o,\mathbf{r}) \propto \int\int\int P(\mathbf{X},\mathbf{Z}|\theta)P(\mathbf{R}|\mathbf{X},\mathbf{Z},\mu)P(\theta|\omega)P(\mu|\eta)d\mu dZ d\mathbf{x}^m$$

Unsupervised Learning - MAR: Models

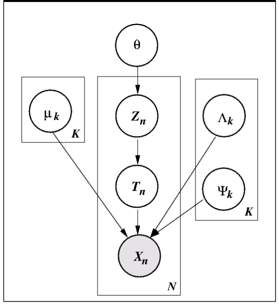
Finite Bayesian Mixture Model



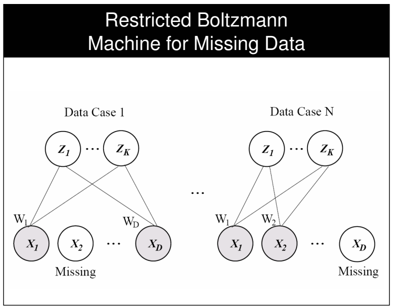
Dirichlet Process Mixture Model

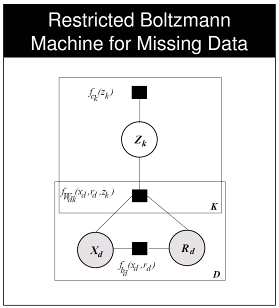


Factor Analysis Mixture Model



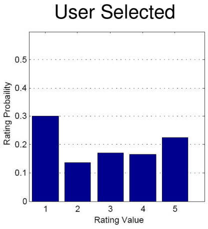
Unsupervised Learning - MAR: Models

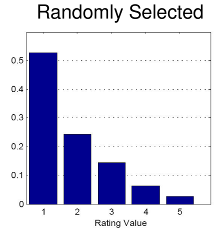




Unsupervised Learning – NMAR: Data Sets: Yahoo!

Collected ratings for randomly selected songs and combined them with existing ratings for user selected songs to form a novel collaborative filtering data set.

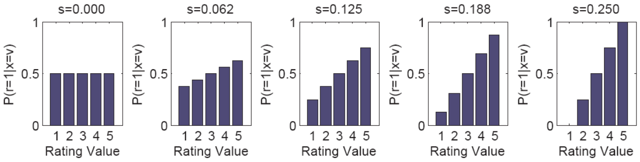




Unsupervised Learning – NMAR: Data Sets: Jester

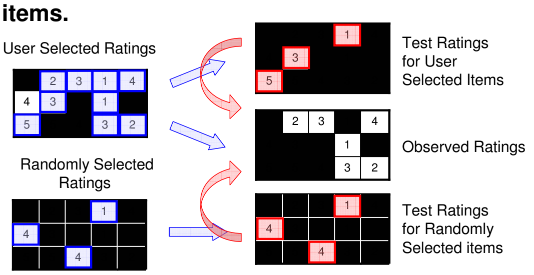
Jester gauge set of 10 jokes used as complete data. Synthetic missing data was added.

- 15,000 users randomly selected
- Missing data model: $\mu_v(s) = s(v-3)+0.5$

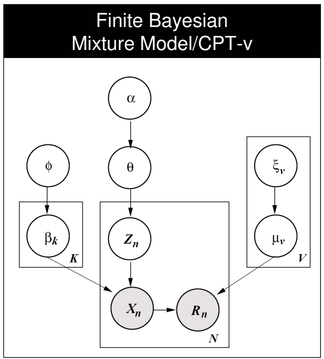


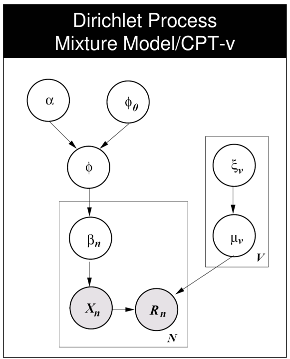
Unsupervised Learning – NMAR: Basic Experimental Protocol

- **We train on ratings for user selected items, and test on ratings for both user selected items, and randomly select items.**

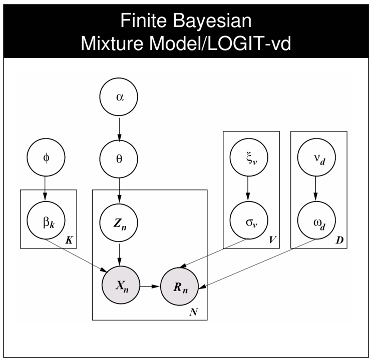


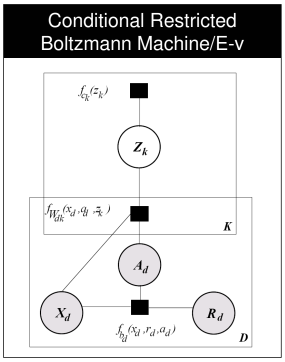
Unsupervised Learning - NMAR: Models





Unsupervised Learning - NMAR: Models





Unsupervised Learning – NMAR: Comparison of Results on Yahoo! Data

	Complexity	Rand MAE	Complexity	User MAE
EM MM	1	0.7725 ± 0.0024	5	0.5779 ± 0.0066
EM MM/CPT-v	20	0.5431 ± 0.0012	10	0.6661 ± 0.0025
EM MM/Logit	5	0.5038 ± 0.0030	5	0.7029 ± 0.0186
EM MM/CPT-v+	5	0.4456 ± 0.0033	20	0.7088 ± 0.0087
MCMC DP	N/A	0.7624 ± 0.0063	N/A	0.5767 ± 0.0077
MCMC DP/CPT-v	N/A	0.5549 ± 0.0026	N/A	0.6670 ± 0.0071
MCMC DP/CPT-v+	N/A	0.4428 ± 0.0027	N/A	0.7537 ± 0.0026
CD RBM	20	0.7179 ± 0.0025	10	0.5513 ± 0.0077
CD cRBM/E-v	1	0.4553 ± 0.0031	20	0.5506 ± 0.0085

Unsupervised Learning – NMAR: NEW: Ranking Results

$$NDCG(n) = \frac{\sum_{i=1}^{T} \frac{2^{x_{ni}^t} - 1}{\log(1+\hat{\pi}(i,n))}}{\sum_{i=1}^{T} \frac{2^{x_{ni}^t} - 1}{\log(1+\pi(i,n))}}$$

- $\hat{x}_{ni}^t$: mean of posterior predictive distribution for test item i.
- $\hat{\pi}(i,n)$: rank of test item i according to $\hat{x}_{ni}^t$.
- $\pi(i,n)$: rank of test item i according to x_{ni}^t .

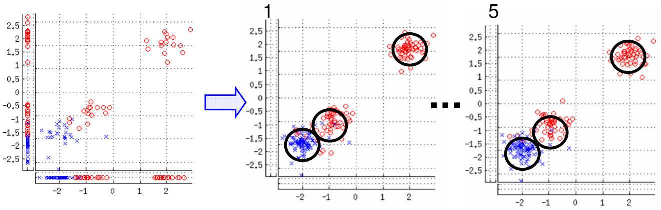
Unsupervised Learning – NMAR: NEW: Comparison of Yahoo! Ranking Results

Strong Generalization:

	Complexity	Rand NDCG
EM MM	1	0.8162 ± 0.0022
EM MM/CPT-v	20	0.8352 ± 0.0023
EM MM/Logit	5	0.8398 ± 0.0012
EM MM/CPT-v+	20	0.8377 ± 0.0012
MCMC DP	N/A	0.8167 ± 0.0025
MCMC DP/CPT-v	N/A	0.8248 ± 0.0020
MCMC DP/CPT-v+	N/A	0.8319 ± 0.0011
CD cRBM	20	0.8207 ± 0.0011
CD cRBM/E-v	10	0.8244 ± 0.0017

Classification: Imputation

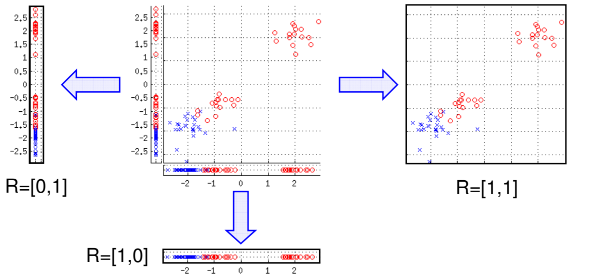
Multiple Imputation: Replace missing feature values with samples of $\mathbf{x}^m$ given $\mathbf{x}^o$ drawn from several imputation models.



Mixture of Factor Analyzers
K=3, Q=1

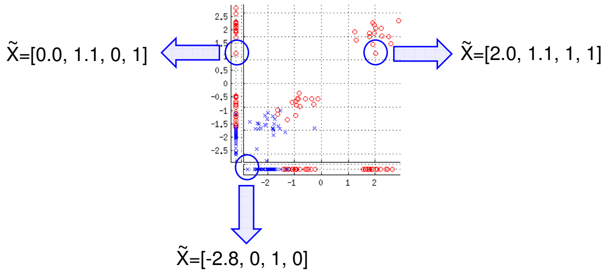
Classification: Reduced Models

Reduced Models: Each observed data subspace defined by a pattern of missing data gives a separate classification problem.



Classification: Response Augmentation

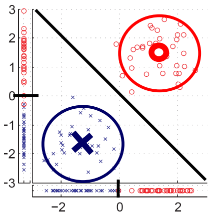
Response Augmentation: Set missing features to zero and augment feature representation with response indicators.



Classification: Generative Models

Generative Model (LDA-FA):

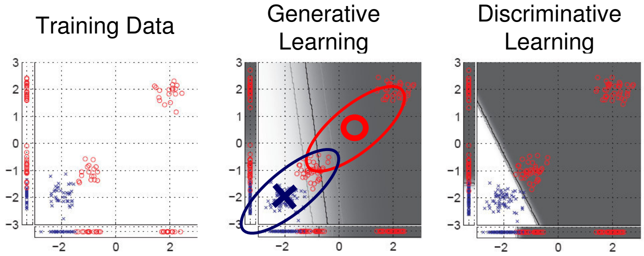
$$P(Y_n = c) = \theta_c$$
$$P(\mathbf{X}_n = \mathbf{x}_n | Y_n = c) = \mathcal{N}(\mathbf{x}_n | \mu_c, \Sigma)$$
$$\Sigma = \Lambda\Lambda^T + \Psi$$



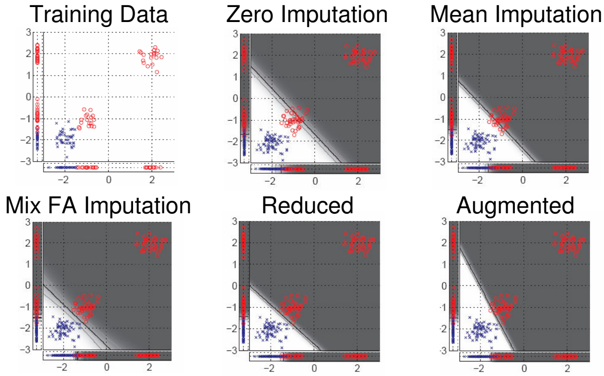
Predictive Distribution with Missing Data:

$$P(Y = c | \mathbf{X}_n^o = \mathbf{x}_n^o) = \frac{\theta_c \mathcal{N}(\mathbf{x}_n^o | \mu_c^o, \Sigma^{oo})}{\sum_c \theta_c \mathcal{N}(\mathbf{x}_n^o | \mu_c^o, \Sigma^{oo})}$$

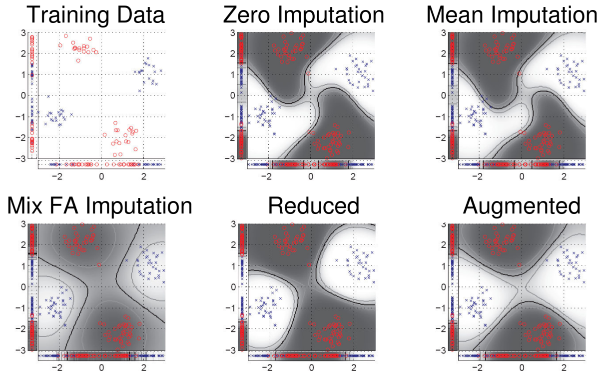
Results: Linear Discriminant Analysis



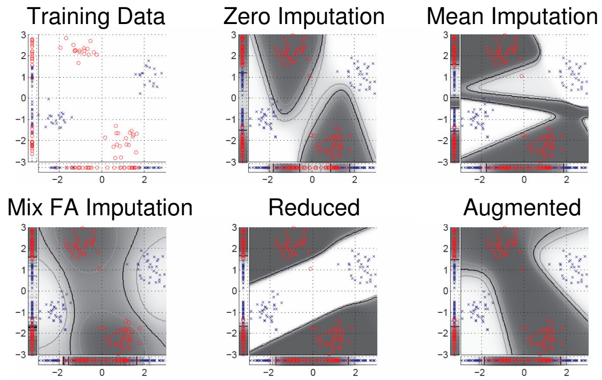
Classification: Logistic Regression



Classification: : Gaussian KLR



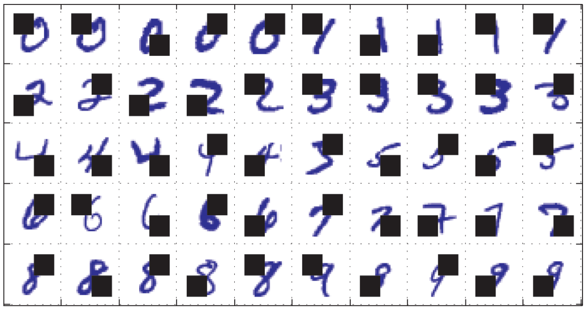
Classification: Neural Networks



Classification:
UCI Thyroid-Sick

	Thyroid: Sick	
	Loss	Err(%)
LR Zero	0.2123 ± 0.0005	6.75 ± 0.00
LR Mean	0.1112 ± 0.0000	5.25 ± 0.00
LR MixFA	0.1270 ± 0.0009	6.21 ± 0.11
LR Reduced	0.1263 ± 0.0000	5.35 ± 0.00
LR Augmented	0.1166 ± 0.0024	5.35 ± 0.06
NN Mean	0.1892 ± 0.0036	6.42 ± 0.00
NN MixFA	0.1118 ± 0.0012	5.03 ± 0.15
NN Reduced	0.1069 ± 0.0022	3.81 ± 0.09
NN Augmented	0.1065 ± 0.0025	4.95 ± 0.19
LDA-FA Dis	0.1092 ± 0.0011	5.16 ± 0.02

Classification:
MNIST Digit Classification with Missing Data



Classification:
MNIST Digit Classification with Missing Data

	MNIST Digits	
	Loss	Err(%)
LR Zero	0.6350 ± 0.0110	19.75 ± 0.41
LR Mean	0.6150 ± 0.0112	19.15 ± 0.34
LR Reduced	0.7182 ± 0.0135	22.62 ± 0.45
LR Augmented	0.6160 ± 0.0112	19.35 ± 0.36
LDA-FA Dis	0.6355 ± 0.0051	19.95 ± 0.25
NN Mean	0.6235 ± 0.0541	18.34 ± 0.42
NN Reduced	0.6944 ± 0.0088	21.51 ± 0.27
NN Augmented	0.5925 ± 0.0161	17.76 ± 0.18
gKLR Mean	0.4147 ± 0.0075	13.02 ± 0.24
gKLR Reduced	0.5694 ± 0.0079	18.32 ± 0.49
gKLR Augmented	0.3896 ± 0.0101	12.34 ± 0.46

The End