Maximum Entropy Density Estimation with Incomplete Presence-Only Data

Bert Huang and Ansaf Salleb-Aouissi {bert@cs, ansaf@ccls}.columbia.edu
Center for Computational Learning Systems, Columbia University, New York, NY 10115

An elegant method to handle missing data when performing density estimation

Learning with Incomplete Data

Common approach:

2	9	1	8	3
?	2	1	5	0
8	?	?	9	4
2	8	5	5	?
3	6	7	?	?

→ Imputation →

2	9	1	8	3
3	2	1	5	0
8	6	3	9	4
2	8	5	5	2
3	6	7	7	2

→ Learning →

0.2
0.1
0.15
0.25
0.3

Only works well if imputation is successful. What if it isn't?

Better approach:

2	9	1	8	3
?	2	1	5	0
8	?	?	9	4
2	8	5	5	?
3	6	7	?	?

→ Learning →

0.2
0.1
0.2
0.2
0.3

Learn only from data we know

Algorithm Derivation

Maximum Entropy: use probability distribution with maximum entropy subject to what is known

Standard Maxent:

$$\max_{p \in \Delta} \quad H(p)$$

$$\text{s.t.} \quad \left| \sum_x p(x) f_j(x) - \tilde{\pi}[f_j] \right| \leq \beta_j$$

Our simple idea: redefine missingness-aware expectation as

$$\frac{\sum_{x \in \text{observed}} p(x) f(x)}{\sum_{x \in \text{observed}} p(x)} = \frac{\sum_x p(x) f_j(x)}{\sum_x p(x) o_j(x)}$$

(i.e., exclude missing values from expectation and renormalize)

New Maxent:

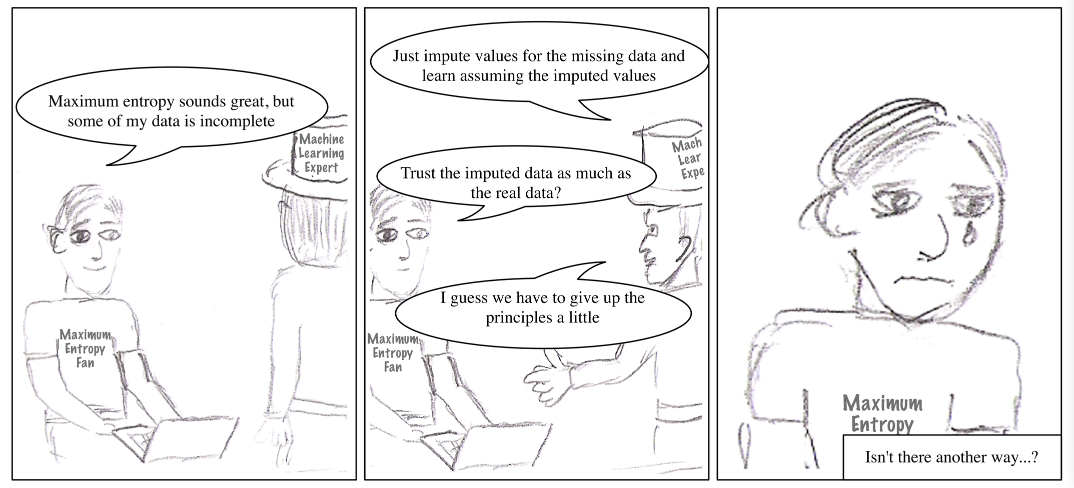
$$\max_{p \in \Delta} \quad H(p)$$

$$\text{s.t.} \quad \left| \frac{\sum_x p(x) f_j(x)}{\sum_x p(x) o_j(x)} - \tilde{\pi}[f_j] \right| \leq \beta_j$$

Dual Formulation $\quad \min_{\lambda} Z(\lambda),$

$$\text{where} \quad p(x) = \frac{1}{Z(\lambda)} \exp \left(\sum_j \lambda_j \left(f_j(x) - \tilde{\pi}[f_k] o_j(x) \right) + |\lambda_j| \, \beta_j o_j(x) \right)$$

Motivation



- Following the Maximum Entropy Principle compels us to avoid assumptions
- We cannot afford to assume we know missing values, even after clever imputation

Missingness and Expectations

The missingness-aware expectation can also be written as: $\sum_x p(x|o_j(x)) f_j(x)$.

I.e., a real expectation over the *naturally occurring* distribution

$$p(x|o_j(x)) = \frac{p(o_j(x)|x)p(x)}{p(o_j(x))},$$

where $p(o_j(x))$ and $p(o_j(x)|x)$ represent the statistical missingness setting.

Imputed expectations inject *artificial* features $g_j(x)$ when data is missing.

$$\sum_x p(x) g_j(x)$$

where $g_j(x) = f_j(x)$ when $o_j(x)$

but $g_j(x) \stackrel{?}{=} f_j(x)$ when $o_j(x)$

Is your $g_j(x)$ good enough?

Notation

Symbol	Definition
$H(p)$	entropy of distribution p
$f_j(x)$	j'th feature for example x
$\tilde{\pi}[f_j]$	empirical average for j'th feature
β_j	(user-defined) allowed expectation deviation
$o_j(x)$	indicator of whether j'th feature is known for example x

Methods Being Compared

Abbreviation	Method
Missing	Our method
Mean (all)	mean imputation to mean of all examples
Mean (positive)	mean imputation to mean of positive examples
EM	Gaussian Expectation-Maximization imputation

Synthetic Data: Missingness Tests

	Missing	Mean (all)	Mean (positive)	EM
MCAR	-261.96 ± 3.04	-262.02 ± 2.99	-262.07 ± 3.05	-262.04 ± 3.05
MAR	-258.58 ± 3.90	-258.70 ± 3.88	-258.75 ± 4.01	-258.63 ± 3.86
NMAR	-258.79 ± 4.02	-259.05 ± 4.01	-258.88 ± 4.22	-259.04 ± 4.00
Full	-254.99 ± 3.30			

- Sampled to simulate missingness settings
- Our approach: highest average out-of-sample likelihood over 5000 synthetic sets

Missingness Settings

Abbreviation	Definition
MCAR	Missing completely at random; missingness is *iid*
MAR	Missing at random; missingness depends on observable features
NMAR	Not missing at random; missingness depends on missing features

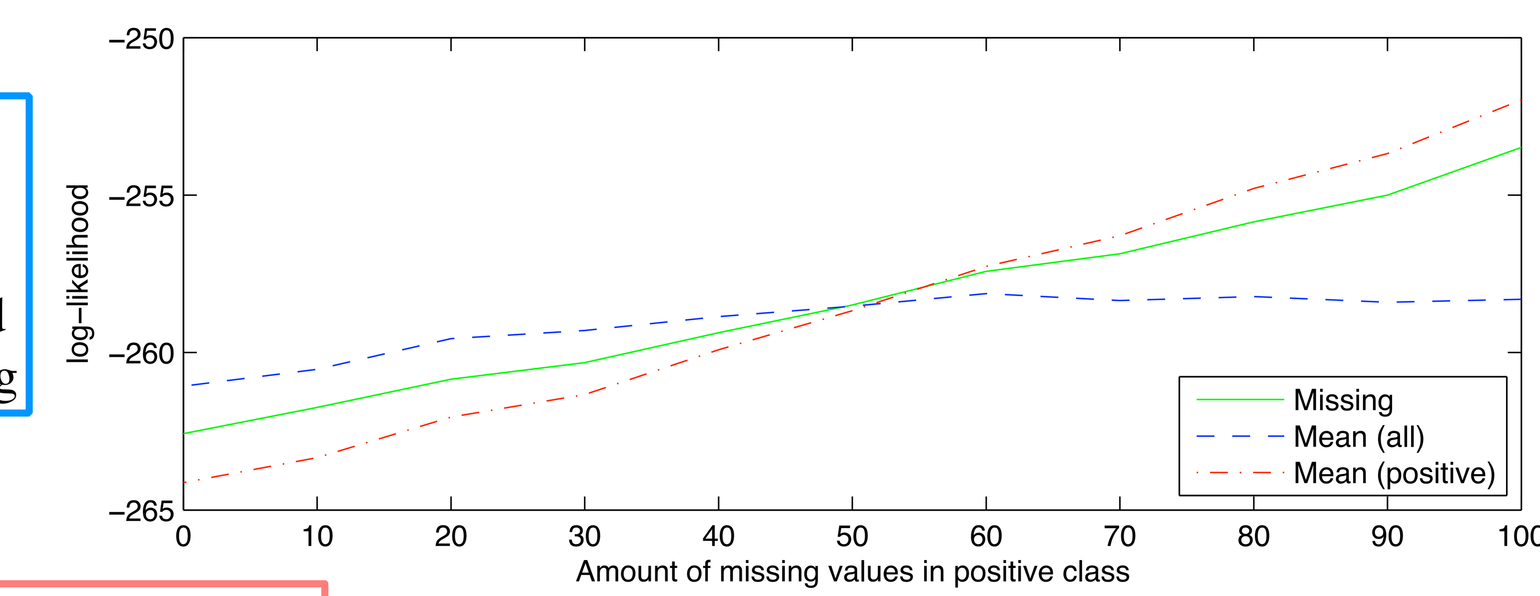
Real Data

	Missing	Mean (all)	Mean (positive)	EM	$p(o)$	$p(y\|o)$
bands	**-711.1±2.9**	−711.5 ± 2.8	**-710.8±2.7**	−711.5 ± 2.8	0.92	0.87
crx	**-991.0±3.2**	**-990.9±3.2**	**-991.0±3.2**	**-990.8±3.3**	0.99	0.42
echo	**-57.2±1.1**	−57.3 ± 1.1	**-57.1±1.0**	−57.4 ± 1.0	0.92	0.32
hep	**-74.8±2.6**	−75.4 ± 2.5	−75.2 ± 2.3	**-75.1±2.5**	0.90	0.27
horse-colic	**-299.3±2.9**	−300.1 ± 2.9	−304.7 ± 2.2	**-299.6±2.7**	0.66	0.38
house-votes	**-555.1±2.8**	**-555.1±2.8**	**-555.1±2.5**	**-555.0±2.8**	0.91	0.42

- UCI data sets with real missing values, run over 500 random training/testing splits
- Chose regularizer via cross-validation, out-of-sample log likelihoods reported in table
- Algorithms use the least complete half of the features (to exacerbate missingness)
- Best and not-statistically worse in **bold** (via 2-sample t-test with %5 rejection)

Synthetic Data: Imputation Quality Tests

When missing data is mostly not positive, Mean (all) more accurately imputes and results in better learning



When missing data is mostly positive, Mean (positive) accurately imputes and results in better learning

...but Mean (all) is not as good because imputation is less accurate

...but Mean (positive) is not as good because imputation is less accurate

Our method is never the worst-case; it never presumes to know the missing values so it won't guess wrong on them!