Conditional Probability Estimation

Marco Cattaneo
School of Mathematics and Physical Sciences
University of Hull

PGM 2016, Lugano, Switzerland
7 September 2016

MLE of conditional probability

- given: a probabilistic model P_θ with unknown θ, past data D, and events E, Q concerning some new (independent) data

MLE of conditional probability

- given: a probabilistic model P_θ with unknown θ, past data D, and events E, Q concerning some new (independent) data

- MLE of $P_\theta(Q \mid E) = P_\theta(Q \mid D \cap E)$:

$$P_{\hat{\theta}_D}(Q \mid E) \qquad \text{with} \qquad \hat{\theta}_D = \arg\max_\theta P_\theta(D) \qquad\qquad (\text{wrong})$$

MLE of conditional probability

- given: a probabilistic model P_θ with unknown θ, past data D, and events E, Q concerning some new (independent) data
- MLE of $P_\theta(Q \mid E) = P_\theta(Q \mid D \cap E)$:

$$P_{\hat{\theta}_D}(Q \mid E) \quad \text{with} \quad \hat{\theta}_D = \arg\max_\theta P_\theta(D) \qquad (\text{wrong})$$

$$P_{\hat{\theta}_{D \cap E}}(Q \mid E) \quad \text{with} \quad \hat{\theta}_{D \cap E} = \arg\max_\theta P_\theta(D \cap E) \qquad (\text{right})$$

MLE of conditional probability

- given: a probabilistic model P_θ with unknown θ, past data D, and events E, Q concerning some new (independent) data

- MLE of $P_\theta(Q \mid E) = P_\theta(Q \mid D \cap E)$:

$$P_{\hat{\theta}_D}(Q \mid E) \quad \text{with} \quad \hat{\theta}_D = \arg\max_\theta P_\theta(D) \qquad \text{(wrong)}$$

$$P_{\hat{\theta}_{D \cap E}}(Q \mid E) \quad \text{with} \quad \hat{\theta}_{D \cap E} = \arg\max_\theta P_\theta(D \cap E) \qquad \text{(right)}$$

- when P_θ is a (generalized) regression model, and E, Q describe predictors and response, respectively, then there is no difference between (right) and (wrong)

MLE of conditional probability

- given: a probabilistic model P_θ with unknown θ, past data D, and events E, Q concerning some new (independent) data

- MLE of $P_\theta(Q \mid E) = P_\theta(Q \mid D \cap E)$:

$$P_{\hat{\theta}_D}(Q \mid E) \quad \text{with} \quad \hat{\theta}_D = \arg\max_\theta P_\theta(D) \qquad \text{(wrong)}$$

$$P_{\hat{\theta}_{D\cap E}}(Q \mid E) \quad \text{with} \quad \hat{\theta}_{D\cap E} = \arg\max_\theta P_\theta(D \cap E) \qquad \text{(right)}$$

- when P_θ is a (generalized) regression model, and E, Q describe predictors and response, respectively, then there is no difference between (right) and (wrong)

- when P_θ is a Bayesian network, D is a training dataset, and E, Q concern some new instances, then the usual MLE is (wrong), and this partially explains the unsatisfactory performance of MLE for Bayesian networks

conditional probability estimation in Bayesian networks

- given: a DAG with vertices $v \in \mathcal{V}$ representing categorical variables X_v, a complete training dataset D with counts $n(\cdot)$, and conjugate Dirichlet priors with parameters $d(\cdot)$

conditional probability estimation in Bayesian networks

- given: a DAG with vertices $v \in \mathcal{V}$ representing categorical variables X_v, a complete training dataset D with counts $n(\cdot)$, and conjugate Dirichlet priors with parameters $d(\cdot)$

- estimates of local probability models:

$$\hat{p}_D(x_v \mid x_{pa(v)}) = \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})} \qquad \text{(ML)}$$

$$\hat{p}_D(x_v \mid x_{pa(v)}) = \frac{n(x_v, x_{pa(v)}) + d(x_v, x_{pa(v)})}{n(x_{pa(v)}) + d(x_{pa(v)})} \qquad \text{(Bayes)}$$

conditional probability estimation in Bayesian networks

- given: a DAG with vertices $v \in \mathcal{V}$ representing categorical variables X_v, a complete training dataset D with counts $n(\cdot)$, and conjugate Dirichlet priors with parameters $d(\cdot)$

- estimates of local probability models:

$$\hat{p}_D(x_v \mid x_{pa(v)}) = \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})} \qquad \text{(ML)}$$

$$\hat{p}_D(x_v \mid x_{pa(v)}) = \frac{n(x_v, x_{pa(v)}) + d(x_v, x_{pa(v)})}{n(x_{pa(v)}) + d(x_{pa(v)})} \qquad \text{(Bayes)}$$

- estimates of probabilities concerning a new instance:

$$\hat{p}_D(x_\mathcal{Q}) = \sum_{x_{\mathcal{V} \setminus \mathcal{Q}}} \prod_{v \in \mathcal{V}} \hat{p}_D(x_v \mid x_{pa(v)}) = \sum_{x_{\mathcal{V} \setminus \mathcal{Q}}} \prod_{v \in \mathcal{V}} \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})} \qquad \text{(ML)}$$

$$\hat{p}_D(x_\mathcal{Q}) = \sum_{x_{\mathcal{V} \setminus \mathcal{Q}}} \prod_{v \in \mathcal{V}} \hat{p}_D(x_v \mid x_{pa(v)}) = \sum_{x_{\mathcal{V} \setminus \mathcal{Q}}} \prod_{v \in \mathcal{V}} \frac{n(x_v, x_{pa(v)}) + d(x_v, x_{pa(v)})}{n(x_{pa(v)}) + d(x_{pa(v)})} \qquad \text{(Bayes)}$$

conditional probability estimation in Bayesian networks

- estimates of conditional probabilities concerning a new instance:

$$\hat{p}_{D,x_{\mathcal{E}}}(x_{\mathcal{Q}} \mid x_{\mathcal{E}}) = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \hat{p}_D(x_v \mid x_{pa(v)})}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \hat{p}_D(x_v \mid x_{pa(v)})}$$

$$= \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})}}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \frac{n(x_v, x_{pa(v)})}{n(x_{pa(v)})}} \qquad \text{(wrong ML)}$$

$$\hat{p}_{D,x_{\mathcal{E}}}(x_{\mathcal{Q}} \mid x_{\mathcal{E}}) = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \hat{p}_D(x_v \mid x_{pa(v)})}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \hat{p}_D(x_v \mid x_{pa(v)})}$$

$$= \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \frac{n(x_v, x_{pa(v)})+d(x_v, x_{pa(v)})}{n(x_{pa(v)})+d(x_{pa(v)})}}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \frac{n(x_v, x_{pa(v)})+d(x_v, x_{pa(v)})}{n(x_{pa(v)})+d(x_{pa(v)})}} \qquad \text{(Bayes)}$$

conditional probability estimation in Bayesian networks

- estimates of conditional probabilities concerning a new instance:

$$\hat{p}_{D,x_\mathcal{E}}(x_\mathcal{Q} \mid x_\mathcal{E}) = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \hat{p}_{D,x_\mathcal{E}}(x_v \mid x_{pa(v)})}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \hat{p}_{D,x_\mathcal{E}}(x_v \mid x_{pa(v)})} = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_v,x_{pa(v)})}{n(x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_{pa(v)})}}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_v,x_{pa(v)})}{n(x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_{pa(v)})}} \quad \text{(ML)}$$

$$\hat{p}_{D,x_\mathcal{E}}(x_\mathcal{Q} \mid x_\mathcal{E}) = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \hat{p}_{D}(x_v \mid x_{pa(v)})}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \hat{p}_{D}(x_v \mid x_{pa(v)})} = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+d(x_v,x_{pa(v)})}{n(x_{pa(v)})+d(x_{pa(v)})}}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+d(x_v,x_{pa(v)})}{n(x_{pa(v)})+d(x_{pa(v)})}} \quad \text{(Bayes)}$$

conditional probability estimation in Bayesian networks

- estimates of conditional probabilities concerning a new instance:

$$\hat{p}_{D,x_\mathcal{E}}(x_\mathcal{Q} \mid x_\mathcal{E}) = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \hat{p}_{D,x_\mathcal{E}}(x_v \mid x_{pa(v)})}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \hat{p}_{D,x_\mathcal{E}}(x_v \mid x_{pa(v)})} = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_v,x_{pa(v)})}{n(x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_{pa(v)})}}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_v,x_{pa(v)})}{n(x_{pa(v)})+\hat{e}_{D,x_\mathcal{E}}(x_{pa(v)})}} \qquad \text{(ML)}$$

$$\hat{p}_{D,x_\mathcal{E}}(x_\mathcal{Q} \mid x_\mathcal{E}) = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \hat{p}_{D}(x_v \mid x_{pa(v)})}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \hat{p}_{D}(x_v \mid x_{pa(v)})} = \frac{\sum_{x_{\mathcal{V}\setminus(\mathcal{Q}\cup\mathcal{E})}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+d(x_v,x_{pa(v)})}{n(x_{pa(v)})+d(x_{pa(v)})}}{\sum_{x_{\mathcal{V}\setminus\mathcal{Q}}} \prod_{v\in\mathcal{V}} \frac{n(x_v,x_{pa(v)})+d(x_v,x_{pa(v)})}{n(x_{pa(v)})+d(x_{pa(v)})}} \qquad \text{(Bayes)}$$

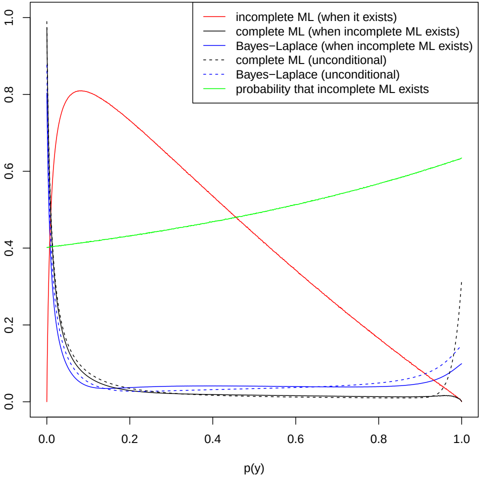
- $\hat{e}_{D,x_\mathcal{E}}(\cdot)$ are the MLE of expected counts for the new instance, obtained from the EM algorithm

performance comparison: $\sqrt{\text{MSE}}$

- given: 3 binary variables X_1, X_2, Y with $X_1 \perp X_2 \mid Y$ and $p(x_1 \mid y) = p(\neg x_1 \mid \neg y) = 99\%$, while $p(\neg x_2 \mid y) = p(\neg x_2 \mid \neg y) = 99\%$

performance comparison: $\sqrt{\text{MSE}}$

- given: 3 binary variables X_1, X_2, Y with $X_1 \perp X_2 \mid Y$ and $p(x_1 \mid y) = p(\neg x_1 \mid \neg y) = 99\%$, while $p(\neg x_2 \mid y) = p(\neg x_2 \mid \neg y) = 99\%$
- estimate $p(y \mid x_1, x_2)$ on the basis of a complete training dataset of size 100:

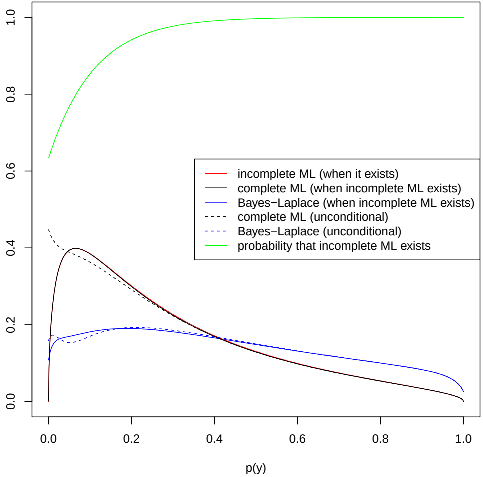


performance comparison: $\sqrt{\mathrm{MSE}}$

- given: 3 binary variables X_1, X_2, Y with $X_1 \perp X_2 \mid Y$ and $p(x_1 \mid y) = p(\neg x_1 \mid \neg y) = 99\%$, while $p(\neg x_2 \mid y) = p(x_2 \mid \neg y) = 90\%$

performance comparison: $\sqrt{\text{MSE}}$

- given: 3 binary variables X_1, X_2, Y with $X_1 \perp X_2 \mid Y$ and $p(x_1 \mid y) = p(\neg x_1 \mid \neg y) = 99\%$, while $p(\neg x_2 \mid y) = p(x_2 \mid \neg y) = 90\%$
- estimate $p(y \mid x_1, x_2)$ on the basis of a complete training dataset of size 100:



conclusion

- the following way of using Bayesian networks is in agreement with Bayes estimation, but not with ML estimation:

  estimate the local probability models of a Bayesian network from data, and then use the resulting global model to calculate conditional probabilities of future events

conclusion

- the following way of using Bayesian networks is in agreement with Bayes estimation, but not with ML estimation:

  estimate the local probability models of a Bayesian network from data, and then use the resulting global model to calculate conditional probabilities of future events

- correct MLE of conditional probabilities can be calculated using the EM algorithm

conclusion

- the following way of using Bayesian networks is in agreement with Bayes estimation, but not with ML estimation:

  estimate the local probability models of a Bayesian network from data, and then use the resulting global model to calculate conditional probabilities of future events

- correct MLE of conditional probabilities can be calculated using the EM algorithm

- future work includes empirical studies of the effect of using the correct MLE on the performance of Bayesian network classifiers