Gaussian kernels for incomplete data

Diego Parente P. Mesquita, Joao Paulo P. Gomes, Francesco Corona, Amauri H. Souza Junior and Juvencio S. Nobre

Department of Computer Science
Federal University of Ceara
Fortaleza, Ceara, Brazil

Today

Summary

- Missing data problem
- Strategies for missing data
- Gaussian Kernels for missing data
- Experiments
- Conclusions

Missing Data Problem

- Observations with one or more missing components, also referred to as incomplete data.
- Standard machine learning methods can not be applied to these data in a straightforward way.
- This might be a problem when the number of missing data is significant.

Strategies for missing data

- Discard samples
- Imputation
  - Nearest Neighbors imputation
  - Expectation Maximization
- Expected Square Distance (Eirola, 2013)
  - For distance based algorithms

Expectation Maximization for missing data

- Iterative process
- Estimates the distribution of the data
- Estimates the expected value of the missing components given the observed ones

Expected Squared Distance

- Estimate the squared distance of two vectors in the presence of missing data
- $X_i, X_j \in \mathbb{R}^D$ drawn from the same multivariate probability distribution, but possibly with deleted entries

$$E[\|X_i - X_j\|_2^2] = \sum_{d=1}^{D} E[(x_{i,d} - x_{j,d})^2]$$

Comparing these two approaches

- Expectation Maximization

$$\|\mathrm{E}[X_i|X_{i,O}] - \mathrm{E}[X_j|X_{j,O}]\|^2$$

$$\mathrm{E}[z] = \sum_{d=1}^{D} (\mathrm{E}[x_{i,d}] - \mathrm{E}[x_{j,d}])^2$$

- Expected Squared Distance

$$\mathrm{E}\left[\|X_i - X_j\|^2 \middle| X_{i,O}, X_{j,O}\right]$$

$$\mathrm{E}[z] = \sum_{d=1}^{D} (\mathrm{E}[x_{i,d}] - \mathrm{E}[x_{j,d}])^2 + \mathrm{var}[x_{i,d}] + \mathrm{var}[x_{j,d}]$$

Gaussian Kernel

- Widely used in machine learning

$$k(X_i, X_j) = \exp\left(-\frac{z_{ij}}{2\sigma^2}\right),$$

- where $z_{ij} = \|X_i - X_j\|^2 = \sum_{d=1}^{D} (x_{i,d} - x_{j,d})^2$

Gaussian Kernel for missing data

Estimate the expected value of the Gaussian Kernel

$$\mathrm{E}[k(z)] = \int_{-\infty}^{+\infty} \mathfrak{p}(z) k(z) dz$$

Gamma distribution for the squared distances (Roberts and Geisser, 1996)

$$\mathfrak{p}(z|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} \exp\left(-\beta z\right),$$

Gaussian Kernel for missing data

Estimate the expected value of the Gaussian Kernel

$$\begin{aligned} \mathrm{E}[k(z)] &= \int_0^\infty \exp\left(-\frac{z}{2\sigma^2}\right)\frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1}\exp\left(-\beta z\right) dz \\ &= M_z\left(-\frac{1}{2\sigma^2}\right) = \left(\frac{2\beta\sigma^2}{2\beta\sigma^2+1}\right)^\alpha \end{aligned}$$

where:

$$\alpha = \frac{\mathrm{E}[z]^2}{\mathrm{var}[z]}, \quad \beta = \frac{\mathrm{E}[z]}{\mathrm{var}[z]}.$$

Gaussian Kernel for missing data

- $\mathrm{E}[z]$ was proposed in (Eirola, 2013)
- For $\mathrm{var}[z]$:

$$\begin{aligned}\mathrm{var}[z] = \Big(\sum_{d=1}^{D} &\mathrm{E}[x_{i,d}^4] + \mathrm{E}[x_{j,d}^4] + 6\mathrm{E}[x_{i,d}^2]\mathrm{E}[x_{j,d}^2] \\ &- 4\mathrm{E}[x_{i,d}]\mathrm{E}[x_{j,d}^3] - 4\mathrm{E}[x_{i,d}^3]\mathrm{E}[x_{j,d}]\Big) \\ + \Big(\sum_{d=1}^{D} \sum_{l \neq d}^{D} &(\mathrm{E}[x_{j,l}^2] + \mathrm{E}[x_{i,l}^2])(\mathrm{E}[x_{j,d}^2] + \mathrm{E}[x_{i,d}^2]) \\ &- 2(\mathrm{E}[x_{i,d}^2] + \mathrm{E}[x_{j,d}^2])\mathrm{E}[x_{i,l}]\mathrm{E}[x_{j,l}] \\ &- 2(\mathrm{E}[x_{i,l}^2] + \mathrm{E}[x_{j,l}^2])\mathrm{E}[x_{i,d}]\mathrm{E}[x_{j,d}] \\ &+ 4\mathrm{E}[x_{i,d}]\mathrm{E}[x_{j,d}]\mathrm{E}[x_{i,l}]\mathrm{E}[x_{j,l}]\Big) - \mathrm{E}[z]^2.\end{aligned}$$

Gaussian Kernel for missing data

- $\mathrm{E}[z]$ and $\mathrm{var}[z]$ are functions of the moments of distribution of the data
- The Expectation Maximization can be used to estimate these moments

Datasets

Computing Gaussian kernels for pairs of instances

Table: Datasets characteristics

Dataset	attributes	instances
FOREST-FIRE (FIRE)	4	517
HABERMAN (HAB)	3	306
DIABETES (PID)	8	768
IRIS	4	150

Results

	%	ICkNNI	ESD	EGK
MPG	10	40.70± 7.13	38.41 ± 4.45	**28.48 ± 3.39**
	50	195.20± 17.81	152.32 ± 5.64	**117.32 ± 6.35**
FIRE	10	451.02± 56.61	340.83 ±15.94	**250.17 ± 16.36**
	50	2050.32± 92.53	1298.47 ± 50.69	**1022.99 ± 52.37**
HAB	10	199.34± 42.71	165.25 ± 18.24	**116.55 ± 14.74**
	50	876.77± 53.80	657.69 ± 16.50	**496.88 ± 18.25**
PID	10	4.05± 1.01	2.40 ± 0.54	**2.08 ± 0.48**
	50	18.50± 1.39	8.54 ± 0.84	**7.80 ± 0.74**
IRIS	10	33.80± 7.24	34.53 ± 8.27	**23.00 ± 5.41**
	50	178.56± 28.44	150.96 ± 14.71	**97.62 ± 8.83**

Conclusions

- We proposed a method (EGK) to calculate the Gaussian kernel on data with missing values
- EGK showed good results in real world data compared to state of the art methods

Thank You!!!