University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

# Bayesian Target Encoding

**Dr. Fani Deligianni,**

**fani.deligianni@glasgow.ac.uk**

**Lecturer (Assistant Professor)**

**Lead of the Computing Technologies for Healthcare Theme**

**https://www.gla.ac.uk/schools/computing/staff/fanideligianni**

WORLD CHANGING GLASGOW

# Bayesian Target Encoding

- In Bayes’ Theorem the probability of a certain event occurring is related to prior knowledge

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- Encode categorical features with moments of the posterior distribution
- Online updating is easy
- Hyperparameters are easy to interpret
- It is easy to generalize the encoding to statistics beyond the mean

# The Beta Distribution

- Bernoulli distribution is suited to model binary target variables (binary classification)
- alpha, beta are the number of positive and negative samples respectively

**PDF**

**Binomial** $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$

**Beta** $\mathrm{g}(p) = \frac{1}{B(a,b)} p^{a-1} (1-p)^{b-1}$

# The Beta Distribution

- Bernoulli distribution is suited to model binary target variables (binary classification)
- alpha, beta are the positive and negative samples respectively

Mean $$\mu = \frac{a}{a+b}$$

Variance $$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

Skewness $$\gamma = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+1)\sqrt{ab}}$$

# Conjugate Bayesian Target Encoding

- Bayes' Theorem about the probability of a certain event occurring is related to prior knowledge

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

- A conjugate Bayesian models assume that the prior and posterior probabilities are of the same distribution family
- Account for variance of posterior distributions
- Account for interactions among other classes

# Summary

- Bayesian Target Encoding is an extension of Mean Target Encoding
- It models higher moments of the posterior distribution of the categories
- It is easy to be updated online
- Parameters are easy to interpret because they relate with the parameters of the underlying probability density functions

# References

- Micci-Barreca. ‘A Pre-processing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems’, ACM SIGKDD Explorations Newsletter 3(1), 2001.
- Slakey et al. Encoding Categorical Variables with Conjugate Bayesian Models for WeWork Lead Scoring Engine’, https://arxiv.org/abs/1904.13001, 2019.