Vector-Quantized-Naive-Bayes

Implementation of the vector-quantized naive Bayes (VQNB) model in Julia.

This is work from UBC's CPSC 540 course, Advanced Machine learning. Note: minor parts of the code (such as the function to compute squared distances, the K-Means implementation, and the sample data file) are from the course, i.e. not created by me. I implemented the vector-quantized naive Bayes method (the VQNB.jl file).

Ordinary naive Bayes is a probabilistic binary classification model. One of it's downfalls is that it does not account for variability within those classes. Take for example the mnist35.jld sample dataset in this repository. It is a collection of labelled, hand-written digits 3 and 5. But there are many different ways people might draw a 3 or a 5. VQNB addresses this by including a latent variable z, which is determined by K-Means clustering. The hope is that these clusterings will represent the "K different ways to draw a 3", and the "K different ways to draw a 5".

In ordinary naive Bayes, we compute the joint probability of the one example as

$$P(x_1^i, \dots, x_d^i, y^i) = P(y^i) \prod_{j=1}^{d} P(x_j^i | y^i) = \theta^{y^i} (1-\theta)^{1-y^i} \prod_{j=1}^{d} \theta_{jy^i}^{x_j^i} (1-\theta_{jy^i})^{1-x_j^i)}$$

for which the parameters $\theta_{\ell c}$ on the right hand side are estimated using the MLE, which works out to be simply

$$\hat{\theta}_{\ell c} = \frac{\sum_{i=1}^{n} x_\ell^i \mathbb{1}_{y^i=c}}{\sum_{i=1}^{n} \mathbb{1}_{y^i=c}}$$

where $\mathbb{1}$ is the indicator function. In vector-quantized naive Bayes, we introduce a latent variable z^i which has a value from 1 to k which is determined by k-means clustering. We compute the joint probability of one example as

$$\begin{aligned} P(x_1^i, \dots, x_d^i, y^i) &= \sum_{z=1}^{k} P(x_1^i, \dots, x_d^i | y^i, z) P(z|y^i) P(y^i) \\ &= P(y^i) \sum_{z=1}^{k} P(z|y^i) \prod_{j=1}^{d} P(x_j^i | y^i, z) \end{aligned}$$

The MLE for $P(x_j^i = 1 | y^i = b, z^i = c)$ is

$$\frac{\sum_{i=1}^{n} x_j^i \mathbb{1}_{y^i=b} \mathbb{1}_{z^i=c}}{\sum_{i=1}^{n} \mathbb{1}_{y^i=b} \mathbb{1}_{z^i=c}}$$

This formula is what determines the `p_xyz` variable in the code for the `VQNB.jl` file.