Relatório do Trabalho Prático de Teoria do Aprendizado Estatístico

Rafael de Sousa Oliveira Martins[1]
NUSP: 10343179

***Abstract*— Trabalho prático da disciplina Teoria do Aprendizado Estatístico.**

I. Objetivo

O objetivo do projeto final é avaliar cinco datasets de diversas áreas e tamanhos. A análise inicial seria do autovalores e autovetores, seguido pela escolha dos melhores parâmetros para o *Support Vector Machine* (SVM), implementado pela libsvm. Após a análise de auto valores e auto vetores, a escolha da melhor combinação de parâmetros para o SVM, será feita a análise do *Genralization Bound (GB)* e o Coeficiente de Shattering.

II. Datasets

Os cincos datasets utilizados para a realização desse projeto foi: ImagetNet Tiny, MNIST, Áudio MSD, PeleSent Buscapé (PLB) e PeleSent MercadoLivre (PLM). Os datasets tiveram seu tamanho e número de classes reduzidos, exceto PLB e PLM. O número de classes de todos datasets são iguais a dois. A ImageNet possuí em seu formato tiny alguns milhares de imagem divididas em 200 classes, utilizamos duas classes selecionada de forma aleatória, e o tamanho do dataset foi de aproximadamente cinco mil imagens, sendo divididas igualmente entre ambas classes. O MNIST de maneira similar foi selecionado duas classes de forma aleatória, contabilizando doze mil imagens de dois dígitos escritos de forma manual. O ES20 possuí áudios de vários gêneros musicais e foi escolhido aleatoriamente dois gêneros, o que contabilizam um base com aproximadamente de mil áudios. Os datasets PLB e PLM, são datasets de textos de reviews de produtos um no buscapé e outro no mercado livre, possuem classes divididas em aproximadamente de mil sentenças de cada classe para cada dataset.

III. Infraestrutura

Todos os dataset utilizados neste projeto são livres e de acesso público para fins de pesquisa. Toda a implementação foi feita utilizando a linguagem de programação Python, em sua versão 3.6, e os algoritmos foram executados utilizando uma máquina com processador i5 3.5ghz, 8gb de memória ram, e sistema operacional Archlinux com kernel 4.14.

IV. Auto vetores e Auto valores

Efetuando a análise de autovetores e autovalores foi possível identificar e mudar o espaço de execução. E a partir da análise dos auto vetores e autovalores foi aplicado o *Principal Components Analysis* (PCA) como mostra a Tabela :

TABLE I
Quantidade de Features antes e depois da análise dos Autovetores e Autovalores

Dataset	Qtd Antes	Qtd Depois
ImageNet	512	198
MNIST	784	182
MSD	30	22
PLB	5013	544
PLM	5013	2589

TABLE II
Quantidade de Features antes e depois da análise dos Autovetores e Autovalores

Dataset	Acurácia	gamma	kernel	coef0
ImageNet	0.94	0.001	polinomial 5	1
MNIST	1.00	0.001	polinomial 3	1
MSD	0.83	0.01	gaussiano	1
PLB	0.75	0.001	polinomial 3	1
PLM	0.59	0.001	polinomial 5	1

De acordo com a tabela , vemos que em todos os casos houve um diminuição de 50% na quantidade features, isso utilizando um valor de 95% na manutenção da variância do dataset. As reduções muito significante como no MNIST ocorre principalmente nod pixels externos, reduzando a imagem à área que realmente é determinante para a classificação, no caso dos dois datasets textuais é devido ao *Bag of Words* que incluem muitas features desnecessárias. No imagenet como as features são os histogramas 3D não possível fazer uma analise de como foi feita essa redução, senão pelo valores dos autovalores e autovetores. Para treinar foi utilizado em todos os casos a validação cruzada em 10 folds e a base foi divida em 75% para treino e 25% para teste.

V. *Support Vector Machines* (SVM)

O SVMs foram executado utilizando a função de *tune* para escolha dentro dos parâmetros selecionados os melhores. Os SVMs foram executados variando o kernel entre polinomial de ordem 2,3,4 e 5, linear e gaussiano. Com estes kernels foram executados utilizando suas versões homogenia e não homogenia, modificando o parametro "coef0"entre 0 e 1. Além da variação na homogeneidade dos kernel variou-se o parâmetro "gamma"em 0.01, 0.001, 0.0001.A avaliação dos melhores parâmetros foi feita utilizando a acurácia. A tabela mostra os melhores parâmetros para cada base com a acurácia obtida.

Como mostra a tabela labeltab2, todos os kernel foram selecionados em sua forma não homogenia. No caso do mnist em especial o resultado de 100% de acurácia foi devido a

TABLE III
QUANTIDADE DE FEATURES ANTES E DEPOIS DA ANÁLISE DOS AUTOVETORES E AUTOVALORES

Dataset	GB
ImageNet	1533.58
MNIST	683.14
MSD	1490.00
PLB	1684.95
PLM	4567.98

seleção aleatorizada escolher os números 1 e 9, números bem distintos entre si.

VI. GENERALIZATION BOUND E SHATTERING COEFFICIENT

O cálculo do *generalization bounde* (GB) foi feita utilizando a fórmula vista em aula, aproximando o "delta"para 0.05, "c"para 4, os demais valores são calculados por cada execução do SVM após a escolha dos melhores parâmetros. A seguir a tabela mostra os valores do GB para cada dataset.

De acordo com a tabela com o calculo do generalization bound não há convergência para essa quantidade de classes e dados, sendo assim, seria necessário mais dados para poder garantir o aprendizado.

VII. CONCLUSÃO

È possível perceber que o svm com os parâmetros corretos obtém excelentes valores de acurácia, mas em contra partida necessita de grandes volumes de dados para algumas tarefas e ainda necessita de uma grande quantidade de processamento de memória, sendo inviável a execução em máquinas de uso pessoal como a utilizadas neste projeto. E apesar de obter uma excelente acurácia em alguns casos e não concluí que foi possuí obter a convergência necessária para se separar a base.

VIII. DIFICULDADES

O projeto teve diversas dificuldades, desde a execução até mesmo a implementação. A dificuldades maior foi executar os datasets, pois no início do prazo foi executado utilizando todo o dataset com suas inúmeras classes, ficando executando durente diversos dias sem obter sequer algum resultado. Devido a isso foi calculado a quantidade de svm's que está sendo feitas, era algo aproximadamente de 320 svm's para o MNIST por exemplo. Sendo assim, houve uma diminuição dos datasets e das classes e mesmo assim ainda causou alguns transtornos como o não cálculo do shaterring coefficient.

Referências utilizadas foram os slides e o livro disponibilizado. Além de matérias das bases disponibilizados em seus sites.