ENGG5103 Techniques for Data Mining
Project Proposal
Survival Prediction On the Titanic

LIN Huangjing
WANG Xi

1. Backgrounds

This is one of the challenges from Kaggle website, with the Challenge Title "*Titanic: Machine Learning from Disaster*"[1].

The sinking of the RMS Titanic which is one of the deadliest peacetime maritime disaster in history happened on the night of 14th April through to the early morning of 15th April in the North Atlantic Ocean on her maiden voyage from Southampton to New York City[2]. The Titanic sank after striking an iceberg, which led to the deaths of more than 5,000 people out of 2224 (The number is right?) passengers and crew.

Mistakes of judgment and operation, delaying warming, improper material of hull and so forth were inferred to result in such loss of life. In addition, there were also deficient lifeboats prepared on the Titanic for the passengers and crew. As a result, over a thousand of passengers and crew were still on board when the Titanic sank and almost of those who had to jump or fall into the water drowned quickly due to the effects of hypothermia. However, in the process of rescue, women, children and the upper-class were given priorities to board the safe refuge, which means those people highly likely survived in the shipwreck.

Lin Huangjing 16/11/5 3:49 PM
已删除: e

In this challenge, we are asked to analyze what category of people tended to survive the tragedy, which would do benefit to make better safety regulations for ships and rescue strategy afterwards.

2. Data and software

The dataset provided includes two parts, one is training data and the other one is testing data[3]. In training data, the information of each person on the Titanic was offered, consisting of PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Far, Cabin and Embarked. The variable descriptions are as followed:

Survived	Suvival (0 = No; 1 = Yes)
Pclass	Passerger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Name	Name
Sex	Sex
Age	Age
SibSp	Number of Siblings/Spouses Aboard
Parch	Number of Parents/Children Aboard
Ticket	Ticket Number

Fare	Passenger Fare
Cabin	Cabin
Embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

There are some samples of training data and testing data respectively below:

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

Fig.1 Training data

	A	B	C	D	E	F	G	H	I	J	K
1	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292		Q
3	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7		S
4	894	2	Myles, Mr. Thomas Francis	male	62	0	0	240276	9.6875		Q
5	895	3	Wirz, Mr. Albert	male	27	0	0	315154	8.6625		S
6	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22	1	1	3101298	12.2875		S
7	897	3	Svensson, Mr. Johan Cervin	male	14	0	0	7538	9.225		S
8	898	3	Connolly, Miss. Kate	female	30	0	0	330972	7.6292		Q
9	899	2	Caldwell, Mr. Albert Francis	male	26	1	1	248738	29		S
10	900	3	Abrahim, Mrs. Joseph (Sophie Halaut Easu)	female	18	0	0	2657	7.2292		C

Fig2. Testing data

Obviously, there is slight difference between training data and testing data. The labels for *Survived* is given in the former but not given in the latter.

As for software, we would like to utilize python for preprocess data and implement the kernel algorithm and analyze the results we get.

3. Objectives

We aim at training a robust binary classifier which could predict the correct category of the unlabeled data from testing dataset. In other words, when one piece of personal information out of testing dataset is given, the classifier could provide with the survival of such person.

Lin Huangjing 16/11/5 3:56 PM
已删除: to

Lin Huangjing 16/11/5 3:57 PM
已删除: training

4. Methods

In this task, we would like to apply *neural network* to such dataset. Deep networks enjoy the great reputation of having significantly excellent representational power. Undeniably,

deep learning methods outperforms other approaches by a large margin in multiple domains currently, such as speech, language, vision and so forth. In addition, it reduces the need for feature engineering which is quite time-consuming in the process of machine learning practice. Besides, sufficient training data including 891 persons' information are provided for us to train our model, therefore overfitting problem could be well avoided. Furthermore, as our model can run on GPU platform, NVIDIA GeForce GTX TITAN X, thus the training time would be considerably cut down. As a consequence, we tend to employ deep neural network method to solve this problem in order to obtain the results with high accuracy efficiently. Not to mention, other approaches such as K-Nearest neighbor algorithm, Random Forest, and Decision Tree will also be used to make comparison with the Deep networks in this task.

5. References

1. https://www.kaggle.com/c/titanic
2. https://en.wikipedia.org/wiki/Sinking_of_the_RMS_Titanic
3. https://www.kaggle.com/c/titanic/data
4. https://www.python.org/