第 1 回春輪講（機械学習コンテスト）

1. 概要

Python を使用した機械学習について，工夫を凝らすことで正答率を上げてください．工夫するために考えたこと，実際の正答率が分かるようなスライドを作成し，春輪講の日に発表してもらいます．

今回の機械学習コンテストでは，
「タイタニック号事件の顧客情報のデータに基づいた生存者予測」
をやっていただきます．

目的は

- 今後の研究で役に立つプログラミング言語「Python」に慣れる
- 分からないことを自分で調べて解決する習慣をつける
- 問題を解決するために考えたアイデアをスライドで伝える能力を身につける

の 3 点です．

2. 作業の例

皆さんに考えていただくことは，大まかに以下の 2 点です．

① 欠損値，文字列の変換
② 特徴の取捨選択

以下ではその詳細について説明します．

① 欠損値，文字列の変換

顧客情報として与えられているデータは以下の通りです．

- PassengerId: kaggle が振っただの連番。

- PassengerClass: 乗客の等級。1 から 3 まで。金持ちが生き残りそう。
- Name: 名前。"Mr."とか"Don."とか色々情報がある。うまくやれば家族の情報も抽出できるかも。
- Sex: 性別。映画では女性、子どもが先に救命ボートに乗ったらしい。関係ありそう。
- Age: 年齢。同上。
- Sibsp: 海外に住んでいる兄弟、配偶者の数。
- Parch: 海外に住んでいる両親、子どもの数。
- Ticket: チケットの番号。"113803"とか"A/5 21171"とか規則がよくわからない。
- Fare: 運賃。金持ちが生き残りそう。
- Cabin: 部屋番号。"C85"みたいな感じ。ぱっと見た感じほとんど欠損している。
- Embarked: 乗船した港。Cherbourg、Queenstown、Southampton の３種類。

トレーニングデータとテストデータにはそれぞれ多数の欠損値（値が全く入っていないセル）や文字列が存在しており，そのままでは数値しか扱うことの出来ない機械学習に落とし込めません．

そこで，みなさんには欠損値を他の値として補完（例．ただ単に全て 0 とする）したり，文字列を別の値に変換（例. male→1, female→0）することで，機械学習に適用できるデータにしてもらいます．

例では全ての欠損値を 0 とすると書きましたが，それでは実際の状況があまりにも再現できず（例．Age の欠損値を 0 にした場合，0 才が多数乗船していることになる），良い正答率は達成できないことが予測できます．このため，この欠損値をどう補完するかが大きなカギの一つとなります．

② 特徴の取捨選択

今回の機械学習では，顧客情報を特徴として扱います．今回は上記の通り全部で 11 つありますが，果たしてその全てが機械学習をする上で有用な特徴と言えるのでしょうか．

例として，Age について考えてみます．実際に災害に遭遇した時，小さな子供は自分の力では対応できませんが，大人になれば筋力・判断力共に向上するため，助かる確率は高いと考えられます．よって，Age を特徴として利用して機械学習をすることは有効と考えられます．

逆に，PassengerId はどうでしょうか．これはただデータの作成者がテキトーに割り振った連番であり，「この数値が高ければ生存率が高い」などということは到底考えられません．

このように，特徴には機械学習に有用なもの，そうでないものが存在しています．皆さんには上記のような理由を考えた上で，機械学習で用いる特徴を取捨選択してもらいます．ただ理由を考えるだけでなく，トレーニングデータ中の特徴と生存の是非をヒストグラム化してみて，有用かどうかを判断してみるのも良いと思います．

以上の 2 点が終われば，後は機械学習です．これは形成し直したデータをプログラムに投げるだけなので，自分で調べてみて実装してみてください． テストデータを入力とした生存の予測のリストを csv ファイルに書き込み，それを kaggle というサイトにアップロードすると以下の式で表される正答率が返ってきます．

$$\text{正答率} = \frac{\text{正解したテストデータ数}}{\text{全テストデータ数}}$$

3. 添付ファイル
   - train.csv－ トレーニングデータです．顧客情報と生存の是非が記載されています．これを学習用データとして利用してください．
   - test.csv－ テストデータです．顧客情報が記載されています．この情報から生存の是非を予測します．
   - gender_submission.csv－ 生存の是非の予測の例です．これを参考にしてcsv ファイルを作成し，kaggle に提出してください．

4. 参考サイト
   - 機械学習をするにあたり，今回は Python の scikit-learn というパッケージを用いることをおすすめします．理由は機械学習が圧倒的に簡単に実装できるからです（トレーニングデータの学習・テストデータの予測ラベルの出力なんかも 1 行で書けます)．Windows に Python やその他必要になりそうなパッケージをインストールする方法が書いてあるページは以下に記載します．
     https://qiita.com/ExA_DEV/items/db7844135eee61a33320
   - 皆さんには生存の是非を予測した csv ファイルを kaggle というサイトにアップロードしていただきます．このためには kaggle に登録する必要がありますので，研究室の G メールアカウント等を利用して各自登録してください．登録後は kaggle にて「Titanic: Machine Learning from Disaster」と検索して該当ページに移動し，画面内の「Submit Predictions」から作成した csv ファイルをアップロードしてもらいます．すぐに結果が返ってきますので，スライドにはそのスクリーンショットを載せてください．kaggle のホームページを以下に記載します．
     https://www.kaggle.com/

5. 最後に
   実装をする上で分からないことがあれば，とりあえず Google 先生に聞きましょう．これも研究の練習です．それでも分からないことがあれは先輩に聞いてください．以上です．頑張って素晴らしい正答率を目指してください．