# PFI/PFN 2015インターン課題

## 提出方法

1. 回答プログラムをメールに添付してください。プログラムの説明や試した手法について書いたものを補足資料として添付してもよいです(ただしA4用紙1枚程度の分量に収めて下さい)。
2. プログラムのビルドに必要な環境、実行環境・実行手順について記述してください。
3. 以下のプログラミング言語のいずれかを利用してください。
    - C, C++, Ruby, Python, Go, Java, Scala, Lua, Cuda, JavaScript

## 評価指標

各問題について、(計算量等は問わず)要件を満たすプログラムを作成できていることが第一です。また、以下の観点でも評価を行います。

- 計算量は少ない方がよい
- メモリ使用量は少ない方がよい
- ユニークなアイディアに基づいているのがよい
- ソースコードの品質は良いのがよい。ただし品質とは何かについては各自考える
- プログラムが正しいかどうかをチェックできている方がよい

## 問題概要

データが以下の形式で与えられます。ただし、label[i] は整数、 data[i, j] は浮動小数点、各データの区切りは半角スペース(0x20)、各行の区切りはCR(0x0A)です。

```
n m
label[1] data[1, 1] data[1, 2] data[1, 3] … data[1, m]
label[2] data[2, 1] data[2, 2] data[2, 3] … data[1, m]
…
label[n] data[n, 1] data[n, 2] data[n, 3] … data[n, m]
```

データは https://www.dropbox.com/s/6z6hwcywxiae6p8/data.zip?dl=0 からダウンロードして下さい。以下のファイルから構成されます。

- `train.txt` : 訓練用データです。課題1,2,3で用いて下さい。
- `train_nolabel.txt` : ラベルの付いていない訓練用データです。課題3.1で使うことが出来ます。
- `test.txt` : テスト用のデータです。課題3.1で訓練したモデルでこのデータを分類して下さい。

以下では2値分類問題を扱います。label[i] は 1 または 2 (ただしラベル無しの場合は 0) です。

1. 課題1 (プログラミング能力を測ります) 標準ライブラリ以外は使わないでください。
   1.1. data の各次元ごとの平均を計算して、data からその平均を引き、各次元のデータセットにわたる平均が0になるように変換する関数を実装し、結果を入力と同じフォーマットで出力するプログラムを実装してください。
   `void SubtractMean(vector<vector<float> >& data);`

   例えば入力が以下のとき
   ```
   2 3
   1 1.0 -9.0 1.0
   2 3.0 -1.0 -1.0
   ```
   以下のような出力をするプログラムを実装せよ
   ```
   2 3
   1 -1.0 -4.0 1.0
   2 1.0 4.0 -1.0
   ```

   1.2. 課題 1.1. の出力 data について、各次元ごとの標本分散を計算して、data の各要素を分散の平方根（標準偏差）で割り、各次元毎の分散を1に変換する関数を実装し、結果を入力と同じフォーマットで出力するプログラムを実装してください。`void NormalizeVariance(vector<vector<float> >& data);`

   例えば入力が以下の時
   ```
   2 3
   1 -1.0 -4.0 1.0
   2 1.0 4.0 -1.0
   ```
   以下のような出力をするプログラムを実装せよ
   ```
   2 3
   1 -1.0 -1.0 1.0
   2 1.0 1.0 -1.0
   ```

1.3. 単純パーセプトロンを用いて学習をして5fold-交差検定でのマクロ平均での正解率 (accuracy) を返す関数を実装してください。正解率は (正しく分類が出来た事例数) / (全事例数) とします。学習と交差検定には `train.txt` を用いて下さい。

```
float CrossValidate(int num_iterations,
                    const vector<vector<float> >& data,
                    cosnt vector<int>& labels);
```

以下に単純パーセプトロンについて簡単に説明します。
m 次元の入力ベクトルを x、ラベルを y とします。ここでは説明しやすくするため、y は+1または -1であるとします。また、m 次元の重みベクトルを w とします。bias 項は無視します。
dot_product(w, x) を w と x の内積とします。この時、単純パーセプトロンは I(dot_product(w, x)) を推定結果とします。ただし、I(z) は z >= 0 の時 +1、z < 0 の時 -1 であるような関数です。
この時、単純パーセプトロンの学習は

```
for num_iterations times do
  for each data x and label y do
    if y * I(dot_product(w, x)) < 0 then
      w ← w + y x
    end if
  end for
end for
```

のように学習します。

2. 課題2 （問題解決能力を測ります）ライブラリを自由に使って良いです。

2.1. 白色化：データの共分散行列が単位行列になるよう変換する関数を実装し、入力データに変換を適用して同じ形式で出力するプログラムを実装してください。

```
void ZCAWhitening(vector<vector<float> >& data);
```

http://deeplearning.stanford.edu/wiki/index.php/Implementing_PCA/Whitening

2.2. 何らかの手法で学習をして、5fold-交差検定で**精度が87%**を超えるようにしましょう。手法が思いつかない場合のお勧めは Averaged Perceptronまたは SVM+RBF カーネルです。

3. 課題3 (オプション) 5fold-交差検定で**精度95%**を達成してください。
※学業に支障が出ない範囲で取り組んでください。

3.1. ラベルがついていない教師なしデータ `train_nolabel.txt` を使えます。入力データ上では label[] の値は0になっています。
訓練済みのモデルでデータ `test.txt` を分類して下さい。分類結果を1つのファイルにまとめ、送付して下さい。ファイルは i 行目に i 番目のデータの分類結果(1 または 2) が書かれた形式にして下さい。