Jubaclassifier Hands-on

主に書籍の内容について、コードを実行しながら説明します

データは書籍とは違うものを使います。

自己紹介

@imaimai1125 (github, twitter)

- Jubatus OSSに2015年11月より参加(ちょうど2年!)
- Jubatus本では**classifier**と**burst**を担当
- 新しい物好き

Agenda

1. 分類とは
2. 実際に動かしてみる

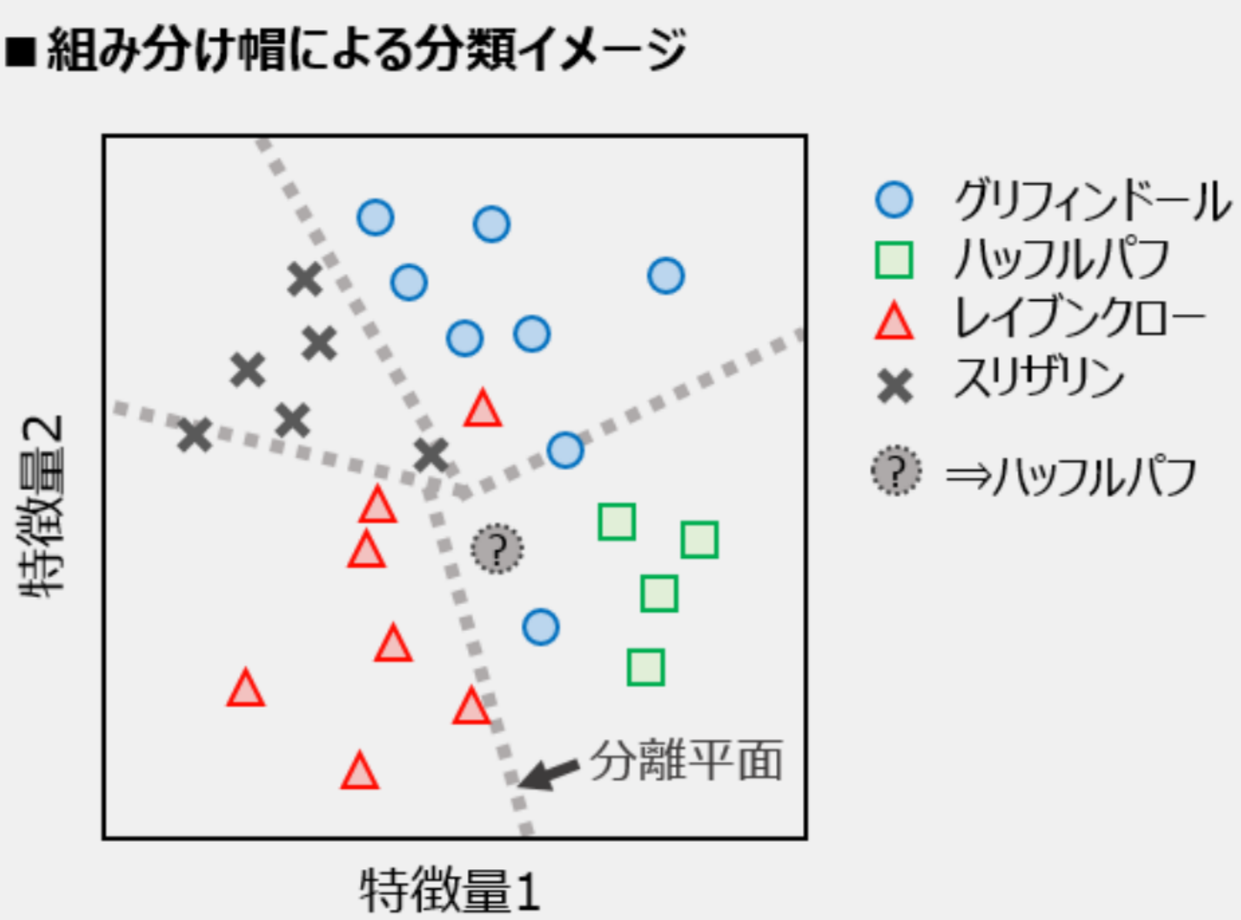
分類とは

分類(Classifier)とは、与えられたデータに対して**適切なクラスを推定する処理**を指します。

- スパムメール判定
- 金融における顧客のデフォルト(不払い)予測
- etc ...

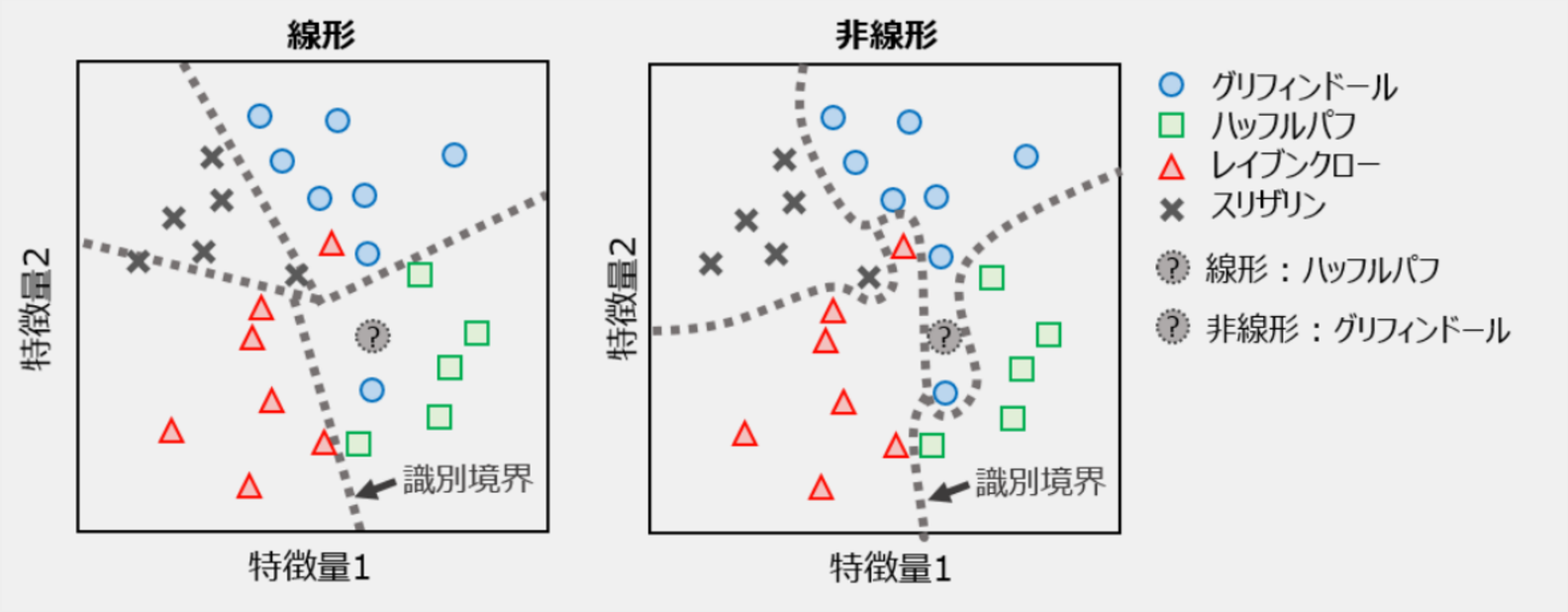
分類のイメージ

Harry Potterの組み分け帽を例にとる



線形分類器 VS 非線形分類器

- 線形分類器：空間を**直線/平面**で区切っていく
  - シンプルで早いが、**どうしても分けられない**場合も出てくる
- 非線形分類器：空間を**曲線/曲面**で区切っていく
  - 精度はいいが、**過学習**の恐れあり



実際に動かしてみる

実際にjubaclassifierを動かすことで、Jubatusについて、分類について学んでいきましょう。

立ち上げ

用意したdockerの人

- 下記コマンドを実施

```
$ docker run -p 8888:8888 -it jubatus/hands-on-5th /bin/bash
$ jupyter notebook &
```

- http://127.0.0.1:8888 にアクセス

VM、すでに環境がある人

- hands-on-5thフォルダまで行き、jupyter notebookを立ち上げる

手順

- Step0. 環境の確認
- Step1. Jubatusの起動
- Step2. データの読み込み、Datumに変換
- Step3. 作成したDatumをJubatus
- Step4. 学習モデルを用いて分類を行う
- Step5. 結果の分析をしてみる
- Step6. 前処理をしてみる
- Step7. 指標のトレードオフ

Step0. 環境の確認

環境に、以下のものが入っているか確認してください。

- Jubatus
- Pythonライブラリ(jubaclassifierハンズオン用)
    - **jupyter, jubatus, scikit-learn**
    - **numpy, pandas, matplotlib**
- Pythonライブラリ(Jubakit, Python pluginハンズオン用)
    - **jubakit, Cython, embedded_jubatus, statsmodels**

配布物

下記のものがフォルダ **hands-on-5th/classifier** 内にあるか確認してください。

- スクリプト(Python Notebookで配布します)
  - hands_on.ipynb
- データ(data/配下)
  - default_train.csv
  - default_test.csv
- コンフィグファイル(config/配下)
  - linear.json (AROW)
  - nonlinear.json (NN/Euclid LSH)

Step1. Jubatusを起動

- ターミナルに戻り、下記コマンドを入力します
- 線形分類器(AROW)を使ってみます

```
$ jubaclassifier -f config/linear.json -t 1000&
```

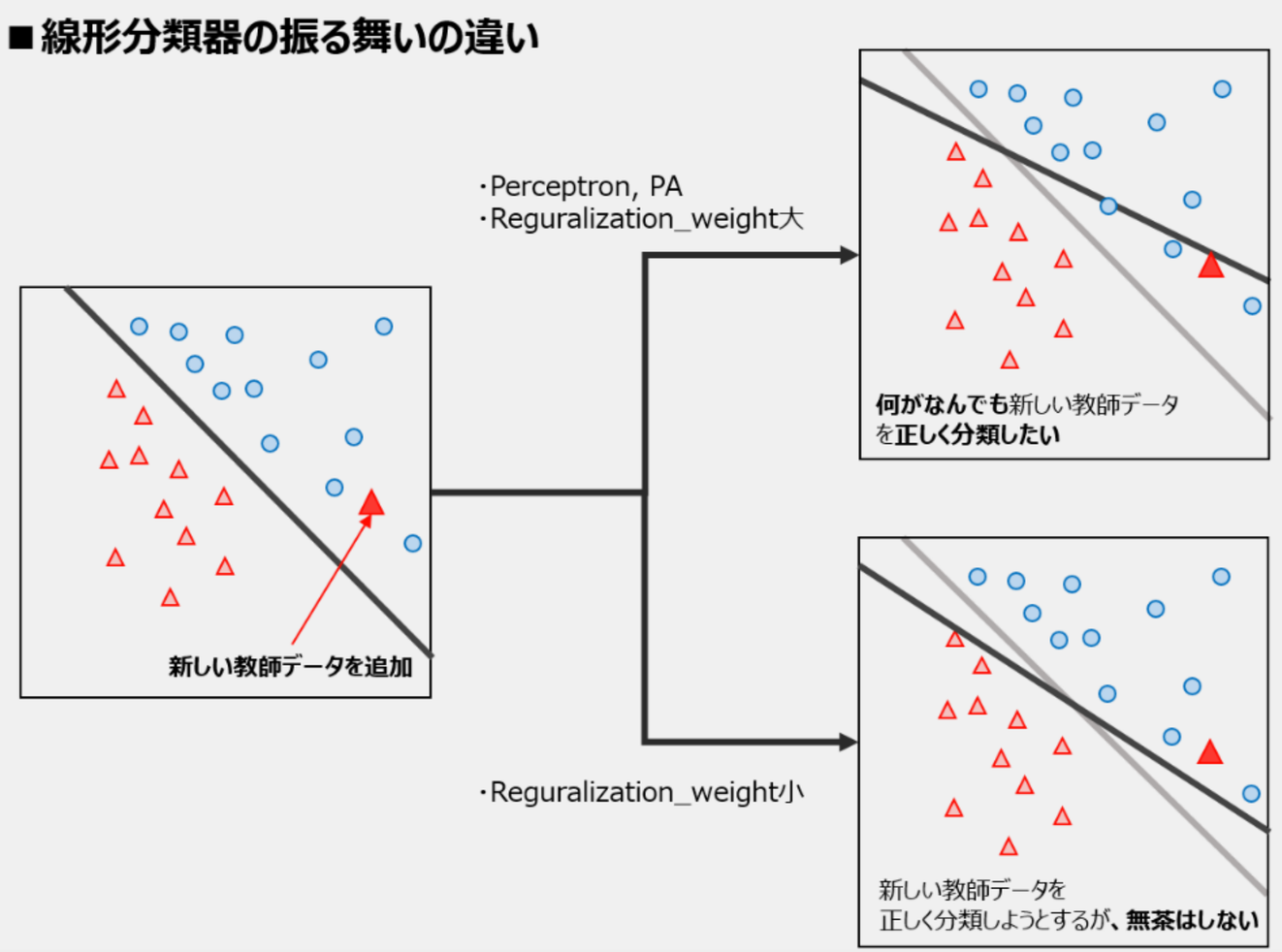
Jubatus分類器に入っているアルゴリズム

Jubatusでは、線形分類器と非線形分類器にそれぞれ下表のアルゴリズムを用意しています。

<table>
<tr><th></th><th colspan="2">アルゴリズム</th></tr>
<tr><td rowspan="6">線形</td><td colspan="2">Perceptron</td></tr>
<tr><td colspan="2">PA</td></tr>
<tr><td colspan="2">PA-I, PA-II</td></tr>
<tr><td colspan="2">CW</td></tr>
<tr><td colspan="2">AROW</td></tr>
<tr><td colspan="2">NHERD</td></tr>
<tr><td rowspan="5">非線形</td><td rowspan="5">k-NN（k近傍法）
それぞれ近傍点の距離の
算出方法が異なる。</td><td>Euclidean</td></tr>
<tr><td>Cosine</td></tr>
<tr><td>euclidLSH</td></tr>
<tr><td>LSH</td></tr>
<tr><td>Minhash</td></tr>
</table>

Regularization Weghtとは

Regularization_weight が大⇛新しく入ってきたデータを何が何でも分類する



Step2. データを読み込み、Datumに変換

- 今回扱うデータ : **default of credit card clients Data Set**
  - 台湾の顧客に対し、支払いの**不払い(デフォルト)があったかどうか**を集めたデータ
  - 年齢や性別など、**23個の特徴量**で構成されている

※本とは違うデータを用いています。

前処理等で、コードが少し異なる箇所もありますが、流れは本のままです。

- **DEFAULT**に支払いが履行されたかどうかの情報が含まれている
- **DEFAULT**を分類器の答え合わせ(正解ラベル)に用いる

列番	特徴量	概要	特徴量の型
X1	Amount of the given credit	信用貸付額	整数
X2	Gender	性別	カテゴリ変数
X3	Education	学歴	カテゴリ変数
X4	Martial status	結婚歴	カテゴリ変数
X5	Age	年齢	整数
X6-X11	History of past payment	過去の支払い。きちんと支払ったかどうか(※1)	整数
X12-X17	Amount of bill statememt	過去の請求額(※1)	整数
X18-X23	Amount of previous payment	過去の支払額(※1)	整数
Y	DEFAULT	デフォルトしたかどうか	カテゴリ変数(正解ラベル)

※1 2005年4月から2005年9月までの6ヶ月分のデータ

Pandasを使ってデータを読み込んでみる

In [2]:
```
import pandas as pd
df = pd.read_csv("data/default_train.csv") #データの読み込み
df.head()
```

Out[2]:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X15	X16	X17	X18	X19	X20	X21	X22	X23	Y
0	20000	2	2	1	24	2	2	-1	-1	-2	...	0	0	0	0	689	0	0	0	0	1
1	120000	2	2	2	26	-1	2	0	0	0	...	3272	3455	3261	0	1000	1000	1000	0	2000	1
2	90000	2	2	2	34	0	0	0	0	0	...	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
3	50000	2	2	1	37	0	0	0	0	0	...	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
4	50000	1	2	1	57	-1	0	-1	0	0	...	20940	19146	19131	2000	36681	10000	9000	689	679	0

5 rows × 24 columns

データの読み込み

- **pandas**を用いてデータを読み込む
- 特徴量ベクトル, ラベル, 特徴量の名前をcsvから出力

データの読み込み

- **pandas**を用いてデータを読み込む
- 特徴量ベクトル, ラベル, 特徴量の名前をcsvから出力

In [3]:
```
def read_dataset(path):
    df = pd.read_csv(path)
    labels = df['Y'].tolist()
    df = df.drop('Y', axis=1)
    features = df.as_matrix().astype(float)
    columns = df.columns.tolist()
    return features, labels, columns

features_train, labels_train, columns = read_dataset("data/default_train.csv")
```

In [4]:

```
print("columns : ¥n{}¥n".format(columns))
print("features : ¥n{}".format(features_train))
```

```
columns :
['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8', 'X9', 'X10', 'X11', 'X12', 'X13', 'X14', 'X15', 'X16', 'X17', 'X18', 'X19', 'X20', 'X21', 'X22', 'X23']

features :
[[  2.00000000e+04   2.00000000e+00   2.00000000e+00 ...,   0.00000000e+00
    0.00000000e+00   0.00000000e+00]
 [  1.20000000e+05   2.00000000e+00   2.00000000e+00 ...,   1.00000000e+03
    0.00000000e+00   2.00000000e+03]
 [  9.00000000e+04   2.00000000e+00   2.00000000e+00 ...,   1.00000000e+03
    1.00000000e+03   5.00000000e+03]
 ...,
 [  1.60000000e+05   1.00000000e+00   6.00000000e+00 ...,   4.50000000e+03
    4.50000000e+03   4.50000000e+03]
 [  8.00000000e+04   1.00000000e+00   2.00000000e+00 ...,   5.52000000e+03
    0.00000000e+00   2.99800000e+03]
 [  2.10000000e+05   1.00000000e+00   3.00000000e+00 ...,   1.00000000e+03
    1.80000000e+03   1.40150000e+04]]
```

学習用データをDatum形式に変換

- Jubatusにデータを投げるために、pandasで読み込んだデータを**Datum形式**にする必要がある
- データと正解ラベルを1セットにしてデータを保持

[(正解ラベル, Datum(key, value))]

In [5]:
```
from jubatus.common import Datum
features_train, labels_train, columns = read_dataset('data/default_train.csv')
train_data = []
for x, y in zip(features_train, labels_train):
    d = Datum({key: float(value) for key, value in zip(columns, x)})
    train_data.append([str(y), d])
```

Step3. 作成したDatumをJubatusサーバに投入

In [6]:
```
from jubatus.classifier.client import Classifier
# Jubatusサーバのホストとポートを指定する
client = Classifier('127.0.0.1', 9199, '')
#過去の学習結果を一度初期化する(任意)
client.clear()
# 学習を実行
client.train(train_data)
```

Out[6]: 24998

Step4. 学習モデルを用いて分類を行う

- 作った学習モデルを使って、実際に分類をしてみる
- **data/default_test.csv**を用いる

※実際の評価の際は、Cross Validationなどがよく使われます。

In [7]:
```
# テスト用Datumリストを作る
features_test, labels_test, columns = read_dataset('data/default_test.csv')
test_data = []
for x, y in zip(features_test, labels_test):
    d = Datum({key: float(value) for key, value in zip(columns, x)})
    test_data.append(d)
```

テストをする

実際にJubaclassifierにデータを投入し、返ってくるラベルを見る

In [8]:
```
# テストをする
results = client.classify(test_data)
```

JubatusがdefaultにたいしてYes/Noどちらが可能性が高いかをスコアリングしてくれている

In [9]:
```
results[0]
```

Out[9]:
```
[estimate_result{label: 1, score: 0.3523704707622528},
 estimate_result{label: 0, score: 0.909548819065094}]
```

Step5. 結果の分析を行う

- 先ほどのスコアの大きい方を分類結果として返すget_most_likely関数を作る
- 結果の混合行列, accuracy, precision, recall, F-valueを算出する

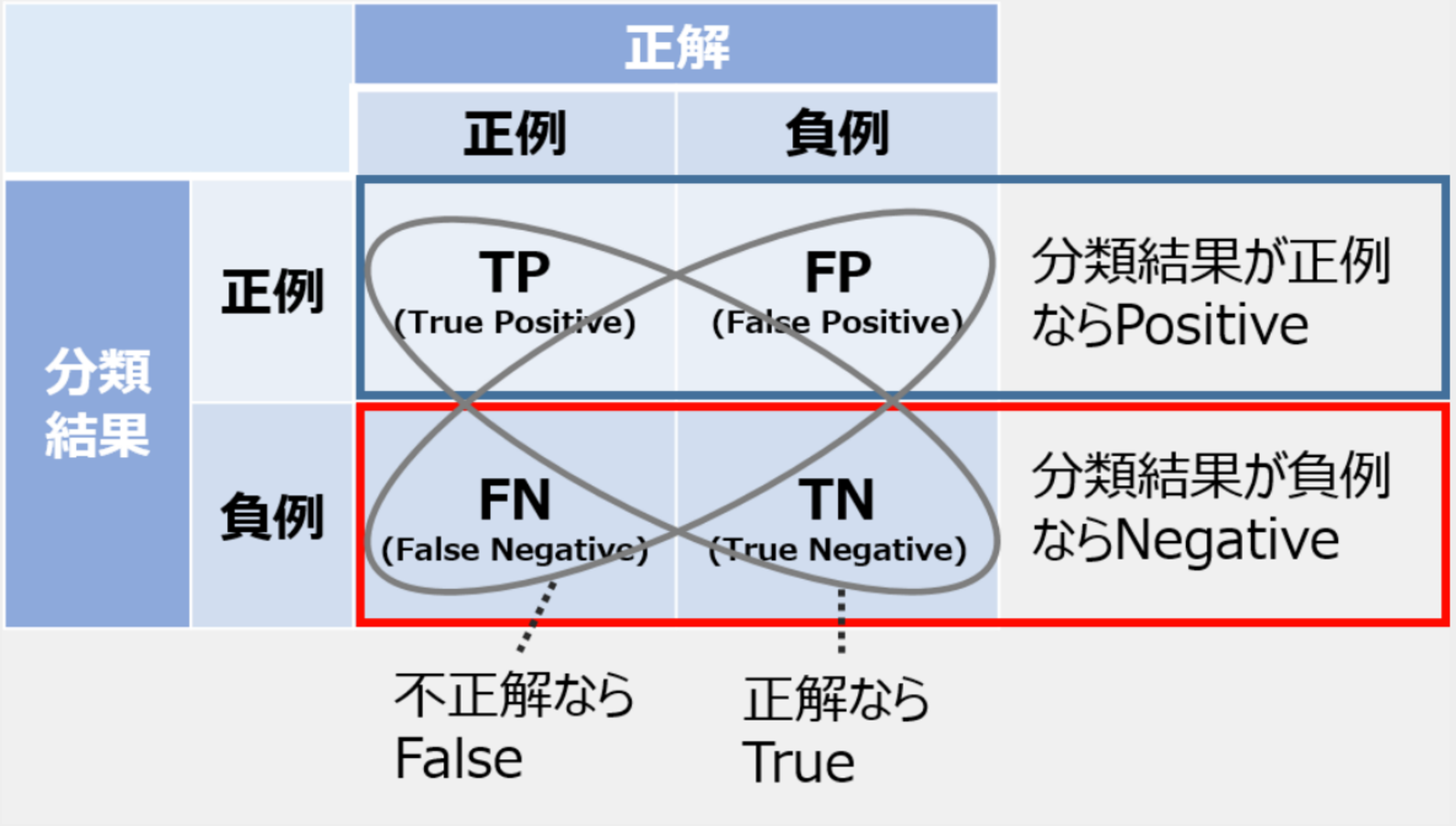
In [10]:
```
# 結果を分析する(スコアの大きい方のラベルを選ぶだけ)
def get_most_likely(result):
    return max(result, key = lambda x: x.score).label
```

分類結果の評価指標

- 分類結果は、
  正例か負例かという観点と、
  答えが合ってたかどうかという観点で4種類に分けられる
- 分類した結果が...
  - Default **有**：正例(**Positive**)
  - Default **無**：負例(**Negative**)
- 推定結果が...
  - 正しい：**True**
  - 間違ってる：**False**

混合行列(Confusion Matrix)

この4種類を一つの表にまとめたもの



精度の評価指標

一口に『**精度**』と言っても、実はいろいろある

Accuracy

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

Precision

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

Recall

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

目的に応じて、見るべき指標は変わる

In [11]:

```
def analyze_results(labels, results, pos_label="1", neg_label="0"):
    tp, fp, tn, fn = 0, 0, 0, 0
    for label, result in zip(labels, results):
        estimated = get_most_likely(result)
        label = str(label)
        estimated = str(estimated)
        if label == pos_label and label == estimated:
            tp += 1
        elif label == pos_label and label != estimated:
            fn += 1
        elif labels != pos_label and label == estimated:
            tn += 1
        else:
            fp += 1
    accuracy = float(tp + tn) / float(tp + tn + fp + fn)
    precision = float(tp) / float(tp + fp)
    recall = float(tp) / float(tp + fn)
    f_value = 2.0 * recall * precision / (recall + precision)
    # confusion matrix
    confusion = pd.DataFrame([[tp, fp], [fn, tn]], index=[pos_label, neg_label], columns=[pos_label, neg_label])
    return confusion, accuracy, precision, recall, f_value
```

In [12]:

```
confusion, accuracy, precision, recall, f_value = analyze_results(labels_test, results)
print('confusion matrix¥n{0}¥n'.format(confusion))
print('metric    : score')
print('Accuracy  : {0:.3f}'.format(accuracy))
print('Precision : {0:.3f}'.format(precision))
print('Recall    : {0:.3f}'.format(recall))
print('F_value   : {0:.3f}'.format(f_value))
```

```
confusion matrix
     1     0
1  308   234
0  751  3708

metric    : score
Accuracy  : 0.803
Precision : 0.568
Recall    : 0.291
F_value   : 0.385
```

Step6. 前処理をしてみる

- Accuracyが高いにも関わらず、**Recallが低い**
- Recall : **Defaultする人に対し、Defaultしたと予測できているか**
    - Recallが低い

      => Defaultするはずの人を捉えられていない

      => リスク管理できない!

Accuracy

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

Precision

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

Recall

		正解	
		正例	負例
分類結果	正例	TP (True Positive)	FP (False Positive)
	負例	FN (False Negative)	TN (True Negative)

Recallを上げるために => アンダーサンプリング

In [13]:
```
print(df[df["Y"]==0].shape)
print(df[df["Y"]==1].shape)
```

```
(19421, 24)
(5577, 24)
```

- 現状
    - 正例(defaultした) : 5577件
    - 負例(defaultしていない) : 19421件
- **負例に引きずられて、分類がうまくいっていない可能性あり**

In [14]:

```python
import random
random.seed(42) # シードで乱数を固定(再現性を得たい場合に実行)
def under_sampling(features, labels, reduce_label, reduce_rate=0.2):
    sampled_features, sampled_labels = [], []
    for feature, label in zip(features, labels):
        label = str(label)
        if label != reduce_label or random.random() < reduce_rate:
            sampled_features.append(feature)
            sampled_labels.append(label)
    return sampled_features, sampled_labels
```

In [15]:

```
# アンダーサンプリング
reduce_rate = 0.2
sampled_features_train, sampled_labels_train = under_sampling(features_train, labels_train,
                                                              reduce_label="0",reduce_rate=reduce_rate)
# 学習用Datumリストを作る
sampled_train_data = []
for x, y in zip(sampled_features_train, sampled_labels_train):
    d = Datum({key: float(value) for key, value in zip(columns, x)})
    sampled_train_data.append([str(y), d])
# 学習をする
client = Classifier('127.0.0.1', 9199, '')
client.clear()
client.train(sampled_train_data)
# テストをする
results = client.classify(test_data)
```

In [16]:

```
# 結果を分析する
confusion, accuracy, precision, recall, f_value = analyze_results(labels_test, results)
print('confusion matrix¥n{0}¥n'.format(confusion))
print('metric    : score')
print('accuracy  : {0:.3f}'.format(accuracy))
print('precision : {0:.3f}'.format(precision))
print('recall    : {0:.3f}'.format(recall))
print('f_value   : {0:.3f}'.format(f_value))
```

```
confusion matrix
     1     0
1  758  1988
0  301  1954

metric    : score
accuracy  : 0.542
precision : 0.276
recall    : 0.716
f_value   : 0.398
```

Step7. 指標のトレードオフ

- PrecisionとRecallはトレードオフの関係にある

metric	score normal	score under sampling
Accuracy	0.803	0.542
Precision	**0.568**	**0.276**
Recall	**0.291**	**0.716**
f_value	0.385	0.398

In [17]:

```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt

rs = np.linspace(0,1.0,11)
precisions = []
recalls = []
```

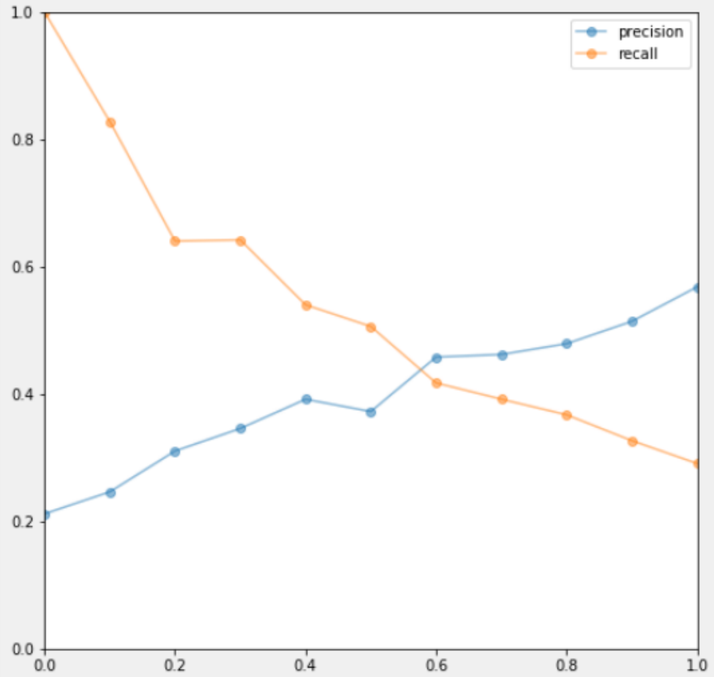
In [18]:

```
for r in rs:
    sampled_features_train, sampled_labels_train = under_sampling(features_train, labels_train, reduce_label="0",reduce_rate=r)
    # 学習用Datumリストを作る
    sampled_train_data = []
    for x, y in zip(sampled_features_train, sampled_labels_train):
        d = Datum({key: float(value) for key, value in zip(columns, x)})
        sampled_train_data.append([str(y), d])
    # 学習をする
    client = Classifier('127.0.0.1', 9199, '')
    client.clear()
    client.train(sampled_train_data)
    # テストをする
    results = client.classify(test_data)
    confusion, accuracy, precision, recall, f_value = analyze_results(labels_test, results)
    precisions.append(precision)
    recalls.append(recall)
#     print(confusion)
    print("rate:{:.1f} precision:{:.2f} recall:{:.2f} accuracy:{:.2f} f-value:{:.2f}".format(r, precision, recall, accuracy, f_v
```

```
rate:0.0 precision:0.21 recall:1.00 accuracy:0.21 f-value:0.35
rate:0.1 precision:0.25 recall:0.83 accuracy:0.43 f-value:0.38
rate:0.2 precision:0.31 recall:0.64 accuracy:0.62 f-value:0.42
rate:0.3 precision:0.35 recall:0.64 accuracy:0.67 f-value:0.45
rate:0.4 precision:0.39 recall:0.54 accuracy:0.73 f-value:0.45
rate:0.5 precision:0.37 recall:0.51 accuracy:0.71 f-value:0.43
rate:0.6 precision:0.46 recall:0.42 accuracy:0.77 f-value:0.44
rate:0.7 precision:0.46 recall:0.39 accuracy:0.77 f-value:0.42
rate:0.8 precision:0.48 recall:0.37 accuracy:0.78 f-value:0.42
rate:0.9 precision:0.51 recall:0.33 accuracy:0.79 f-value:0.40
rate:1.0 precision:0.57 recall:0.29 accuracy:0.80 f-value:0.38
```

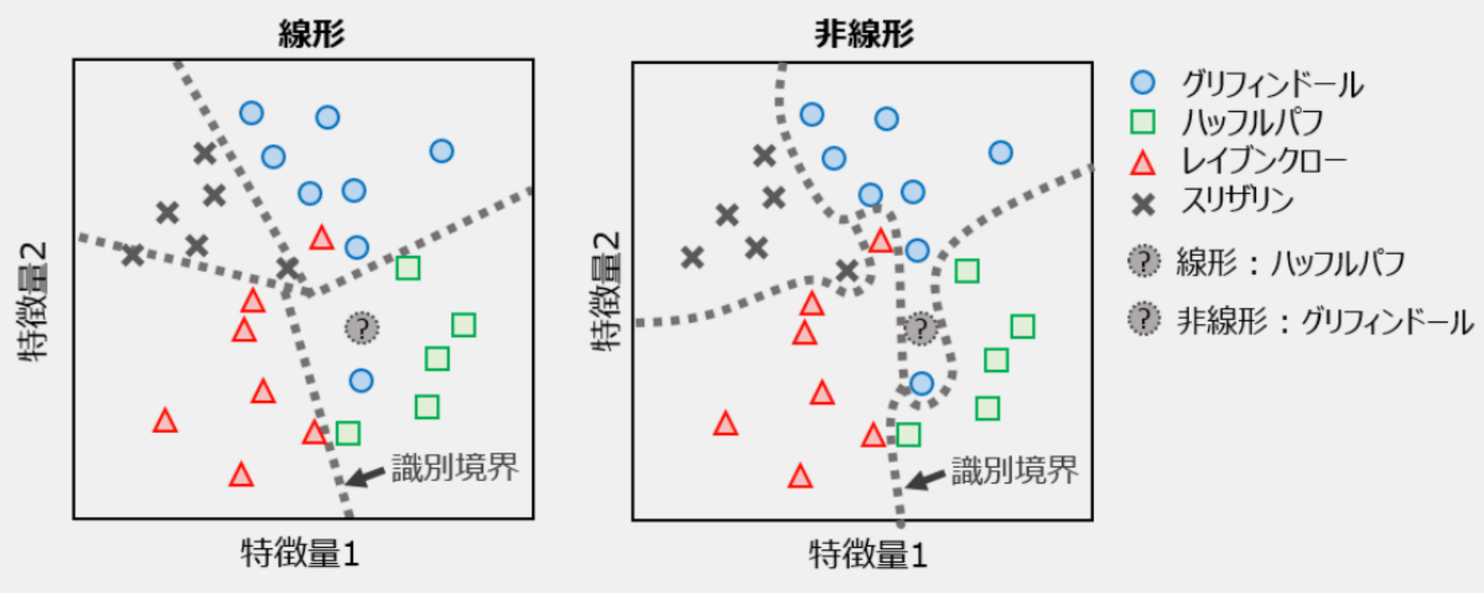
In [19]:

```
plt.figure(figsize=(8,8))
plt.plot(rs, precisions,"o-",alpha=0.5,label="precision")
plt.plot(rs,recalls,"o-",alpha=0.5,label="recall")
plt.xlim(0,1)
plt.ylim(0,1)
plt.legend()
plt.show()
```



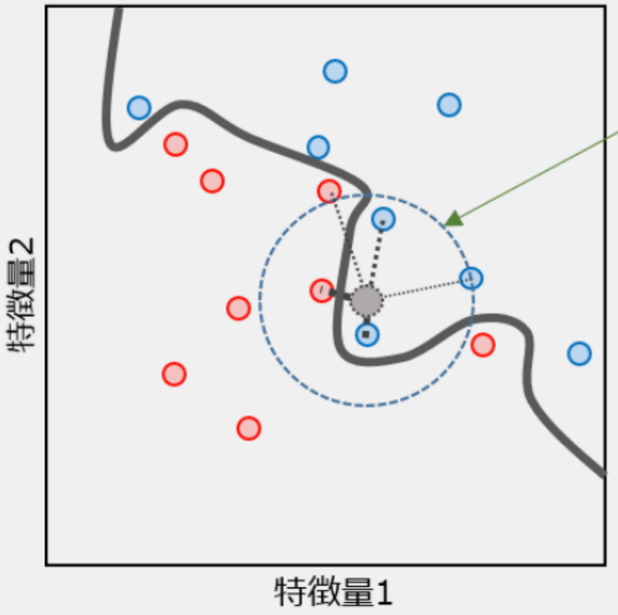
非線形分類器

※今回、非線形分類器の実行は割愛



非線形分類器のパラメータ

■k近傍法(k Nearest Neighbor)



近傍点をk個(図は$k = 5$)を見て、
自分がどっちに行くかを決める
近いほど影響力大

クラス1　クラス2

新しい点がどっちのクラスなのかは、
スコアで決める

$Score(\text{クラス}) = \exp(-\text{distance} * S)$
Sは、距離に対する感度

設定可能な変数
近傍点の数：k　距離に対する感度：S　距離の測り方：**distance**

まとめ

- Jubatusを使って分類をやってみた
- 分類の評価指標はたくさんあるので、**目的に合わせて選ぶ**
- データをうまく処理することで、線形分類器でも精度を上げることが可能

NEXT : Jubakit

より簡単に分析が可能に!