8E and 8F: Finding the Probability P(Y==1|X)

8E: Implementing Decision Function of SVM RBF Kernel ¶

After we train a kernel SVM model, we will be getting support vectors and their corresponsing coefficients α_i Check the documentation for better understanding of these attributes: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)

Attributes:

support_ : *array-like, shape = [n_SV]*
Indices of support vectors.

support_vectors_ : *array-like, shape = [n_SV, n_features]*
Support vectors.

n_support_ : *array-like, dtype=int32, shape = [n_class]*
Number of support vectors for each class.

dual_coef_ : *array, shape = [n_class-1, n_SV]*
Coefficients of the support vector in the decision function. For multiclass, coefficient for all 1-vs-1 classifiers. The layout of the coefficients in the multiclass case is somewhat non-trivial. See the section about multi-class classification in the SVM section of the User Guide for details.

coef_ : *array, shape = [n_class * (n_class-1) / 2, n_features]*
Weights assigned to the features (coefficients in the primal problem). This is only available in the case of a linear kernel.

coef_ is a readonly property derived from `dual_coef_` and `support_vectors_`.

intercept_ : *array, shape = [n_class * (n_class-1) / 2]*
Constants in decision function.

fit_status_ : *int*
0 if correctly fitted, 1 otherwise (will raise warning)

probA_ : *array, shape = [n_class * (n_class-1) / 2]*

probB_ : *array, shape = [n_class * (n_class-1) / 2]*
If probability=True, the parameters learned in Platt scaling to produce probability estimates from decision values. If probability=False, an empty array. Platt scaling uses the logistic function `1 / (1 + exp(decision_value * probA_ + probB_))` where `probA_` and `probB_` are learned from the dataset [R20c70293ef72-2]. For more information on the multiclass case and training procedure see section 8 of [R20c70293ef72-1].

As a part of this assignment you will be implementing the `decision_function()` of kernel SVM, here decision_function() means based on the value return by `decision_function()` model will classify the data point either as positive or negative Ex 1: In logistic regression After traning the models with the optimal weights w we get, we will find the value $\frac{1}{1+\exp(-(wx+b))}$, if this value comes out to be < 0.5 we will mark it as negative class, else its positive class Ex 2: In Linear SVM After traning the models with the optimal weights w we get, we will find the value of $sign(wx + b)$, if this value comes out to be -ve we will mark it as negative class, else its positive class. Similarly in Kernel SVM After traning the models with the coefficients α_i we get, we will find the value of $sign(\sum_{i=1}^{n}(y_i \alpha_i K(x_i, x_q)) + intercept)$, here $K(x_i, x_q)$ is the RBF kernel. If this value comes out to be -ve we will mark x_q as negative class, else its positive class. RBF kernel is defined as: $K(x_i, x_q)$ = $exp(-\gamma ||x_i - x_q||^2)$ For better understanding check this link: https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation (https://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation)

Task E

> 1. Split the data into X_{train}(60), X_{cv}(20), X_{test}(20)
> 2. Train $SVC(gamma = 0.001, C = 100.)$ on the (X_{train}, y_{train})
> 3. Get the decision boundry values f_{cv} on the X_{cv} data i.e. f_{cv} = `decision_function(`X_{cv}`)` you need to implement this decision_function()

In [1]:

```
import numpy as np
import pandas as pd
from sklearn.datasets import make_classification
import numpy as np
from sklearn.svm import SVC
```

In [2]:

```
from sklearn.model_selection import train_test_split
X, y = make_classification(n_samples=5000, n_features=5, n_redundant=2,
                           n_classes=2, weights=[0.7], class_sep=0.7, random_state=15)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y)
X_train, X_cv, y_train, y_cv = train_test_split(X_train, y_train, test_size=0.25, stratify=
X_train.shape, X_cv.shape, X_test.shape
```

Out[2]:

```
((3000, 5), (1000, 5), (1000, 5))
```

Pseudo code

clf = SVC(gamma=0.001, C=100.)
clf.fit(Xtrain, ytrain)
def decision_function(Xcv, ...): #use appropriate parameters
    for a data point x_q in Xcv:
        #write code to implement $(\sum_{i=1}^{\text{all the support vectors}}(y_i \alpha_i K(x_i, x_q)) + intercept)$, here the values y_i, α_i, and $intercept$ can be obtained from the trained model
return # *the decision_function output for all the data points in the Xcv*

fcv = decision_function(Xcv, ...) # *based on your requirement you can pass any other parameters*

Note: Make sure the values you get as fcv, should be equal to outputs of clf.decision_function(Xcv)

In [4]:

```
# you can write your code here
from sklearn.svm import SVC
import math

def decision_function(X_cv,clf):
  cv_desc_func = []
  for x_q in X_cv:
        decision_x_q = np.sum(clf.dual_coef_*np.exp(-clf._gamma*np.sum((clf.support_vectors
        cv_desc_func.append(decision_x_q)
  return np.array(cv_desc_func)

gamma_val = 0.001
clf = SVC(C=100, kernel='rbf', gamma=gamma_val)
clf.fit(X_train,y_train)
decis_func = clf.decision_function(X_cv)
print(decis_func[:5])
decis_func_calc = decision_function(X_cv,clf)
print(decis_func_calc[:5])
```

```
[-0.74439228  1.77714849  1.64888727  1.52845213  0.55574548]
[[-0.74439228]
 [ 1.77714849]
 [ 1.64888727]
 [ 1.52845213]
 [ 0.55574548]]
```

8F: Implementing Platt Scaling to find P(Y==1|X)

Check this PDF (https://drive.google.com/open?id=133odBinMOIVb_rh_GQxxsyMRyW-Zts7a)

Let the output of a learning method be $f(x)$. To get calibrated probabilities, pass the output through a sigmoid:

$$P(y=1|f) = \frac{1}{1+exp(Af+B)} \tag{1}$$

where the parameters A and B are fitted using maximum likelihood estimation from a fitting training set (f_i, y_i). Gradient descent is used to find A and B such that they are the solution to:

$$\underset{A,B}{argmin}\{-\sum_i y_i log(p_i) + (1-y_i)log(1-p_i)\}, \tag{2}$$

where

$$p_i = \frac{1}{1+exp(Af_i+B)} \tag{3}$$

Two questions arise: where does the sigmoid train set come from? and how to avoid overfitting to this training set?

If we use the same data set that was used to train the model we want to calibrate, we introduce unwanted bias. For example, if the model learns to discriminate the train set perfectly and orders all the negative examples before the positive examples, then the sigmoid transformation will output just a 0,1 function. So we need to use an independent calibration set in order to get good posterior probabilities. This, however, is not a draw back, since the same set can be used for model and parameter selection.

To avoid overfitting to the sigmoid train set, an out-of-sample model is used. If there are N_+ positive examples and N_- negative examples in the train set, for each training example Platt Calibration uses target values y_+ and y_- (instead of 1 and 0, respectively), where

$$y_+ = \frac{N_+ + 1}{N_+ + 2};\ y_- = \frac{1}{N_- + 2} \tag{4}$$

For a more detailed treatment, and a justification of these particular target values see (Platt, 1999).

TASK F

1. Apply SGD algorithm with (f_{cv}, y_{cv}) and find the weight W intercept b `Note: here our data is of one dimensional so we will have a one dimensional weight vector i.e W.shape (1,)`

Note1: Don't forget to change the values of y_{cv} as mentioned in the above image. you will calculate y+, y- based on data points in train data

Note2: the Sklearn's SGD algorithm doesn't support the real valued outputs, you need to use the code that was done in the `'Logistic Regression with SGD and L2'` Assignment after modifying loss function, and use same parameters that used in that assignment.

```
def log_loss(w, b, X, Y):
    N = len(X)
    sum_log = 0
    for i in range(N):
        sum_log += Y[i]*np.log10(sig(w, X[i], b)) + (1-Y[i])*np.log10(1-sig(w, X[i], b))
    return -1*sum_log/N
```

if Y[i] is 1, it will be replaced with y+ value else it will replaced with y- value

1. For a given data point from X_{test}, $P(Y = 1|X) = \frac{1}{1+exp(-(W*f_{test}+b))}$ where f_{test} = `decision_function(`X_{test}`)`, W and b will be learned as metioned in the above step

In [4]:

```
from sklearn import linear_model
import math
import matplotlib.pyplot as plt
from tqdm import tqdm

def initialize_weights(dim):
    ''' In this function, we will initialize our weights and bias'''
    #initialize the weights to zeros array of (1,dim) dimensions
    #you use zeros_like function to initialize zero, check this link https://docs.scipy.org
    #initialize bias to zero

    w = np.zeros_like(dim)
    b = 0

    return w,b

def sigmoid(z):
    ''' In this function, we will return sigmoid of z'''
    # compute sigmoid(z) and return
    sigmoid_z = 1/(1+math.exp(-1*z))
    return sigmoid_z

def gradient_dw(x,y,w,b,alpha,N):
    '''In this function, we will compute the gardient w.r.to w '''
    dw = x*(y-sigmoid(np.matmul(w.transpose(),x)+b))-(alpha/N)*w
    return dw

def gradient_db(x,y,w,b):
     '''In this function, we will compute gradient w.r.to b '''
     db = y-sigmoid(np.matmul(w.transpose(),x)+b)
     return db

def pred(w,b, X):
    N = len(X)
    predict = []
    for i in range(N):
        z=np.dot(w,X[i])+b
        if sigmoid(z) >= 0.5: # sigmoid(w,x,b) returns 1/(1+exp(-(dot(x,w)+b)))
            predict.append(1)
        else:
            predict.append(0)
    return np.array(predict)

def logloss(y_true,y_pred):
    '''In this function, we will compute log loss '''
    temp = 0
    for i in range(len(y_true)):
      y_pred_float = y_pred.astype(np.float)
      if y_pred_float[i]<=0.0 or y_pred_float[i]==1.0:
        if y_pred_float[i]==0.0:
          y_pred_float[i] = 1e-12
        elif y_pred_float[i]==1.0:
          y_pred_float[i] = 1-1e-12
        else:
          y_pred_float[i] = abs(y_pred_float[i])
      temp = temp+((y_true[i]*math.log10(y_pred_float[i]))+((1-y_true[i])*math.log10(1-y_pr
    loss = -1*(temp/len(y_true))
    return loss
```

```
def train(X_train,y_train,X_test,y_test,epochs,alpha,eta0):
    ''' In this function, we will implement logistic regression'''
    #Here eta0 is learning rate
    #implement the code as follows
    # initalize the weights (call the initialize_weights(X_train[0]) function)
    # for every epoch
        # for every data point(X_train,y_train)
           #compute gradient w.r.to w (call the gradient_dw() function)
           #compute gradient w.r.to b (call the gradient_db() function)
           #update w, b
        # predict the output of x_train[for all data points in X_train] using w,b
        #compute the loss between predicted and actual values (call the loss function)
        # store all the train loss values in a list
        # predict the output of x_test[for all data points in X_test] using w,b
        #compute the loss between predicted and actual values (call the loss function)
        # store all the test loss values in a list
        # you can also compare previous loss and current loss, if loss is not updating then
    w,b=initialize_weights(X_train[0])
    train_loss = []
    test_loss = []
    for j in tqdm(range(epochs)):
      for i in range(len(X_train)):
        dw = gradient_dw(X_train[i],y_train[i],w,b,alpha,len(X_train))
        db = gradient_db(X_train[i],y_train[i],w,b)
        w = w+eta0*dw
        b = b+eta0*db
      y_true_train = y_train
      y_pred_train = pred(w,b, X_train)
      loss = logloss(y_true_train,y_pred_train)
      train_loss.append(loss)
      y_true_test = y_test
      y_pred_test = pred(w,b, X_test)
      loss = logloss(y_true_test,y_pred_test)
      test_loss.append(loss)
    plt.plot(list(range(epochs)),train_loss,label="Train dataset loss")
    plt.plot(list(range(epochs)),test_loss,label="Test dataset loss")
    plt.xlabel("epochs")
    plt.ylabel("Loss")
    plt.legend()
    plt.show()
    return w,b

num_pos=0
num_neg=0
for i in range(len(y_train)):
    if y_train[i]>0:
        num_pos+=1
    else:
        num_neg+=1
y_pos = (num_pos+1)/(num_pos+2)
y_neg = 1/(num_neg+2)
y_cv = np.where(y_cv==1,y_pos,y_neg)

clf = linear_model.SGDClassifier(eta0=0.0001, alpha=0.0001, loss='log', random_state=15, pe
clf.fit(X_train, y_train)
f_cv = clf.decision_function(X_cv)
f_test = clf.decision_function(X_test)

alpha=0.0001
eta0=0.0001
N=len(f_cv)
```

```
epochs=14
w,b=train(f_cv,y_cv,f_test,y_test,epochs,alpha,eta0)
print(w)
print(b)
```

```
-- Epoch 1
Norm: 0.11, NNZs: 5, Bias: -0.055714, T: 3000, Avg. loss: 0.665312
Total training time: 0.00 seconds.
-- Epoch 2
Norm: 0.22, NNZs: 5, Bias: -0.104952, T: 6000, Avg. loss: 0.615582
Total training time: 0.00 seconds.
-- Epoch 3
Norm: 0.32, NNZs: 5, Bias: -0.148501, T: 9000, Avg. loss: 0.573666
Total training time: 0.00 seconds.
-- Epoch 4
Norm: 0.41, NNZs: 5, Bias: -0.187073, T: 12000, Avg. loss: 0.538114
Total training time: 0.00 seconds.
-- Epoch 5
Norm: 0.50, NNZs: 5, Bias: -0.221359, T: 15000, Avg. loss: 0.507724
Total training time: 0.00 seconds.
-- Epoch 6
Norm: 0.58, NNZs: 5, Bias: -0.251881, T: 18000, Avg. loss: 0.481548
Total training time: 0.00 seconds.
-- Epoch 7
```

Note: in the above algorithm, the steps 2, 4 might need hyper parameter tuning, To reduce the complexity of the assignment we are excluding the hyerparameter tuning part, but intrested students can try that

If any one wants to try other calibration algorithm istonic regression also please check these tutorials

1. http://fa.bianp.net/blog/tag/scikit-learn.html#fn:1 (http://fa.bianp.net/blog/tag/scikit-learn.html#fn:1)
2. https://drive.google.com/open?id=1MzmA7QaP58RDzocB0RBmRiWfl7Co_VJ7 (https://drive.google.com/open?id=1MzmA7QaP58RDzocB0RBmRiWfl7Co_VJ7)
3. https://drive.google.com/open?id=133odBinMOIVb_rh_GQxxsyMRyW-Zts7a (https://drive.google.com/open?id=133odBinMOIVb_rh_GQxxsyMRyW-Zts7a)
4. https://stat.fandom.com/wiki/Isotonic_regression#Pool_Adjacent_Violators_Algorithm (https://stat.fandom.com/wiki/Isotonic_regression#Pool_Adjacent_Violators_Algorithm)