

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 corresponding 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:	<p>support_ : array-like, shape = [n_SV] Indices of support vectors.</p> <p>support_vectors_ : array-like, shape = [n_SV, n_features] Support vectors.</p> <p>n_support_ : array-like, dtype=int32, shape = [n_class] Number of support vectors for each class.</p> <p>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.</p> <p>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.</p> <p>coef_ is a readonly property derived from dual_coef_ and support_vectors_.</p> <p>intercept_ : array, shape = [n_class * (n_class-1) / 2] Constants in decision function.</p> <p>fit_status_ : int 0 if correctly fitted, 1 otherwise (will raise warning)</p> <p>probA_ : 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 $1 / (1 + \exp(\text{decision_value} * \text{probA_} + \text{probB_}))$ 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].</p>
-------------	---

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 training 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 training the models with the optimal weights w we get, we will find the value of $\text{sign}(wx + b)$, if this value comes out to be -ve we will mark it as negative class, else its positive class.

Similarly in Kernel SVM After training the models with the coefficients α_i we get, we will find the value of $\text{sign}(\sum_{i=1}^n (y_i \alpha_i K(x_i, x_q)) + \text{intercept})$, 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 $SVC(\text{gamma} = 0.001, C = 100.)$ on the (X_{train}, y_{train})

3. Get the decision boundary values f_{cv} on the X_{cv} data i.e. $f_{cv} = \text{decision_function}(X_{cv})$ **you need to implement this decision_function()**

```
In [4]: import numpy as np
import pandas as pd
from sklearn.datasets import make_classification
import numpy as np
from sklearn.svm import SVC
```

```
In [5]: 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=42)
```

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  $(\sum_{i=1}^{\text{all the support vectors}} (y_i \alpha_i K(x_i, x_q)) + \text{intercept})$ , 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 [6]: from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.20, stratify=y)
x_train, x_cv, y_train, y_cv = train_test_split(x_train, y_train, test_size=0.25,
```

```
In [7]: x_train.shape
```

```
Out[7]: (3000, 5)
```

```
In [8]: x_cv.shape
```

```
Out[8]: (1000, 5)
```

```
In [9]: x_test.shape
```

```
Out[9]: (1000, 5)
```

```
In [10]: clf = SVC(gamma = 0.001, C = 100)
clf.fit(x_train,y_train)
```

```
Out[10]: SVC(C=100, gamma=0.001)
```

```
In [11]: dec = clf.decision_function(x_cv)
```

```
In [12]: def decision_func(clf,xcv):
    decision = []
    for x_q in xcv:
        dec = np.sum(clf.dual_coef_ * np.exp(-clf.gamma*np.sum((clf.support_vectors_ - x_q)**2)))
        decision.append(dec)
    return np.array(decision)
```

```
In [13]: dec[0:5]
```

```
Out[13]: array([-0.14328323, -2.21472069, -4.42832628,  1.87987879, -2.46758273])
```

```
In [14]: f_cv = decision_func(clf,x_cv)
```

```
In [32]: f_cv[0:5]
```

```
Out[32]: array([[ -0.14328323],
                [-2.21472069],
                [-4.42832628],
                [ 1.87987879],
                [-2.46758273]])
```

First 5 values of both sklearn's decision function and custom decision function match as shown

above

```
In [29]: f_test = decision_func(clf,x_test)
```

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

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

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

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:

$$\operatorname{argmin}_{A,B} \left\{ - \sum_i y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right\}, \quad (2)$$

where

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

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

$$y_+ = \frac{N_+ + 1}{N_+ + 2}; \quad y_- = \frac{1}{N_- + 2} \quad (4)$$

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

TASK F

4. 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 be replaced with y_- value

5. 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 mentioned in the above step

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 hyperparameter tuning part, but intrested students can try that

If any one wants to try other calibration algorithm isotonic 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=1MzmA7QaP58RDzocB0RBmRiWfI7Co_VJ7 (https://drive.google.com/open?id=1MzmA7QaP58RDzocB0RBmRiWfI7Co_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)

```
In [33]: N_plus = np.count_nonzero(y_train)
N_minus = len(y_train) - N_plus
y_plus = (N_plus + 1) / (N_plus + 2)
y_minus = 1 / (N_minus + 2)
y_plat_cv = [y_plus if i > 0 else y_minus for i in y_cv]
```

```
In [34]: 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
    #initialize bias to zero

    w = np.zeros_like(dim)
    b = 0
    return w,b
```

```
In [35]: def sigmoid(z):
    sig = 1 / (1+np.exp(-z))
    return sig
```

```
In [36]: def logloss(y_true,y_pred):
    n = len(y_true)
    loss = 0
    for i in range(n):
        loss = loss + (y_true[i] * (np.log10(y_pred[i]))) + ((1-y_true[i]) * np.l
    loss = - 1 * (loss / n)
    return loss
```

```
In [37]: 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.dot(np.transpose(w),x) + b)) - ((alpha / N)*w)
    return dw
```

```
In [38]: def gradient_db(x,y,w,b):
    '''In this function, we will compute gradient w.r.to b '''
    db = y - sigmoid(np.dot(np.transpose(w),x) + b)
    return db
```

```
In [39]: import pdb
def Sgd_algo(X_train,y_train,epochs,alpha,eta0):
    ''' In this function, we will implement logistic regression'''
    N = len(X_train)
    w,b = initialize_weights(X_train[0])
    train_loss = []
    test_loss = []
    for i in range(epochs):
        for j in range(N):
            dw = gradient_dw(X_train[j],y_train[j],w,b,alpha,N)
            db = gradient_db(X_train[j],y_train[j],w,b)
            w = w + (eta0*dw)
            b = b + (eta0*db)
            z = np.dot(X_train,w) + b
            y_pred_tr = sigmoid(z)
            tr_loss = logloss(y_train,y_pred_tr)
            train_loss.append(tr_loss)

        if i > 3:
            if train_loss[-1] == train_loss[-2]:
                break

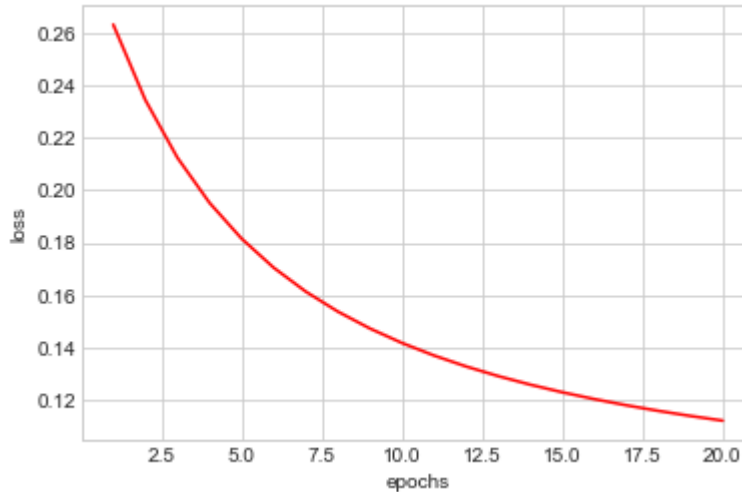
    return w,b,train_loss,
```

```
In [47]: alpha=0.0001
eta0=0.0001
epochs=20
w,b,loss=Sgd_algo(f_cv,y_plat_cv,epochs,alpha,eta0)
```

```
In [49]: %matplotlib inline
import matplotlib.pyplot as plt

plt.style.use('seaborn-whitegrid')
fig = plt.figure()
ax = plt.axes()
epoch = [i for i in range(1,21)]
ax.plot(epoch,loss,color='red')
plt.xlabel('epochs')
plt.ylabel('loss')
```

Out[49]: Text(0, 0.5, 'loss')



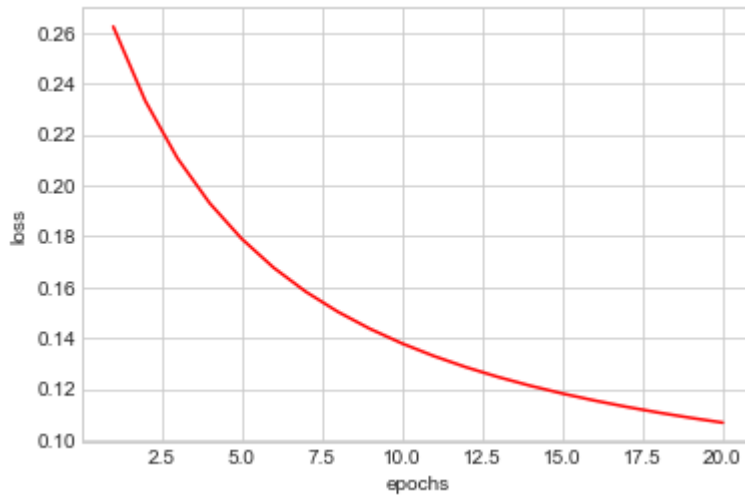
```
In [41]: w2,b2,loss2 =Sgd_algo(f_test,y_test,epochs,alpha,eta0)
```

```
In [44]: prob = 1/(1+np.exp(-(w2*f_test+b2)))
```



```
In [50]: plt.style.use('seaborn-whitegrid')
fig = plt.figure()
ax = plt.axes()
epoch = [i for i in range(1,21)]
ax.plot(epoch,loss2,color='red')
plt.xlabel('epochs')
plt.ylabel('loss')
```

Out[50]: Text(0, 0.5, 'loss')



In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

In []:

