CS 156a Final

Amitesh Pandey

December 2024

Problem 1

Let's take a point (x, y) of dimension $d = 2$ and first apply a polynomial transform of dimension $Q = 2$.

$$(x, y) = (x, y, xy, x^2, y^2) = (x^1y^0, x^0y^1, x^1y^1, x^2y^0, x^0y^2)$$

We can think of this as a counting problem where we must count all possible permutations of the exponents of x and y such that their sum does not exceed Q. For $Q = 10$, note that none of the two exponents can exceed 10 since it isn't permissible to have the other exponent negative. Also observe that if we pick α as the exponent for x, then we $11 - \alpha$ choices for the exponent of y, say β. Now α itself can range from 0 to 10. So the dimension d'

$$d' = \sum_{\alpha=0}^{10}(11 - \alpha) = 66$$

We will have a transform of the form

$$(x, y) \to (1, x, x^2, x^3, \ldots, x^5y^5, \ldots, y^{10})$$

Notice that this does include the constant coordinate 1 produced by the exponent arrangement x^0y^0, ignoring this, we get a total dimension of 65. So the correct option is **[e]**.

Problem 2

For this problem,

[a] If $\mathcal{H}$ has only one hypothesis, say h, then for all $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$, we have $g^{(\mathcal{D})} = h$. This implies that $\bar{g} = \frac{1}{n} \cdot (h + h + \cdots + h) = h$. But we know that $h \in \mathcal{H}$.

[b] Say for $\mathcal{D}_k$, the received hypothesis is $g^{(\mathcal{D}_k)} = a_k$. Then $\bar{g} = \frac{1}{n}(a_1 + \cdots + a_n)$. This is a real value, so $\bar{g} \in \mathcal{H}$.

[c] Say for $\mathcal{D}_k$, the received hypothesis is $g^{(\mathcal{D}_k)} = a_kx + b_k$. Then $\bar{g} = \frac{1}{n}(a_1x + b_1 + \cdots + a_nx + b_n)$. On refactorising, we get $\bar{g} = \frac{1}{n}(a_1 + \cdots + a_n)x + \frac{1}{n}(b_1 + \cdots + b_n)$. Both are real values, so we have $\bar{g} \in \mathcal{H}$.

[d] Say for $\mathcal{D}_k$, the received hypothesis $g^{(\mathcal{D}_k)} = \frac{1}{1+e^{-(a_k+b_kx)}}$. Then

$$\bar{g} = \frac{k + e^{-(a_1+b_1x)} + e^{-(a_2+b_2x)} + \cdots + e^{-(a_n+b_nx)}}{(1 + e^{-(a_1+b_1x)})(1 + e^{-(a_2+b_2x)}) \ldots (1 + e^{-(a_n+b_nx)})}$$

Clearly $\bar{g} \notin \mathcal{H}$, so the correct option is **[d]**.

Problem 3

Suppose there are two hypotheses h, h'. By definition, h' overfits relative to h if it achieves a lower in-sample-error, say E'_{in}, but a higher out-sample-error, say E'_{out} than h. Assume h achieves in-sample-error E_{in} and out-sample-error E_{out}. Naturally, $E_{\text{in}} \neq E'_{\text{in}}$ and $E_{\text{out}} \neq E'_{\text{out}}$. It follows then that $E'_{\text{out}} - E'_{\text{in}} \neq E_{\text{out}} - E_{\text{in}}$. More specifically, $E'_{\text{out}} - E'_{\text{in}} > E_{\text{out}} - E_{\text{in}}$. This inequality makes it clear that overfitting detection requires some sort of a *baseline* $E_{\text{out}} - E_{\text{in}}$ level to which a hypothesis' results need to be compared. This means that overfitting cannot be determined with only one hypothesis. This eliminates options **[a]**, **[b]**, **[c]**, **[e]**. The main problem with comparing only $E_{\text{out}} - E_{\text{in}}$ values is that generally, a hypothesis overfits relative to another not only if the difference in its out-sample and in-sample errors are large, but also that this large difference is necessarily produced by a large E_{out} and a relatively small E_{in}, or that the hypothesis has *overfit* the data. This detail is lost in comparing only the difference value. So the correct option is **[d]**.

Problem 4

Deterministic noise refers to the component of noise borne out of a foundational difference in the complexity of the data and the capacity of the hypothesis set. *Stochastic noise* refers to more *random* noise that's a result of inaccurate measurements, random processes intercepting the data-generation method etc. With this in mind

[a] Stochastic and deterministic noise do not contradict each other. The processes that they originate from can coexist. So this statement is false.

[b] Since deterministic noise is in some sense a reflection of the hypothesis sets' incapacity, we can conclude that it does depend on the hypothesis sets. So this statement is false.

[c] Since deterministic noise is in some sense a reflection of the target function's complexity, we can conclude that it does depend on the target function. So this statement is false.

[d] Since stochastic noise deals more with randomness in the "input" side of the learning diagram, I would imagine it is not effected by our choice of hypotheses. So this is true.

[e] As stated previously, stochastic noise is a measure of the randomness in the distribution of $\mathcal{X}$ and the randomness of the target function. So this statement is false.

In conclusion, option **[d]** is correct.

Problem 5

Recall that the unconstrained solution to the regularization problem is $\mathbf{w}_{\text{lin}}$. Now we're given in the question that this vector satisfies the Tikhonov constraint. Then, this is also the constrained solution. So $\mathbf{w}_{\text{reg}} = \mathbf{w}_{\text{lin}}$, so **[a]**.

Problem 6

Soft order constraints that regularise polynomials can be used to regularise the hypothesis, in this case, they are directly associated with the augmented error. So **[b]** is the correct option.

Problem 7

Using the code in the appendix, with only the bias $(x_1, x_2) \to (1, x_1, x_2)$, we get that the "8 versus all" classifier has the smallest E_{in} with $E_{\text{in}} \approx 0.74$. The correct option is **[d]**.

Problem 8

Using the code in the appendix, with the transform $(x_1, x_2) \to (x_1, x_2, x_1x_2, x_1^2, x_2^2)$, we get that the "1 versus all" classifier has the smallest E_{out} with $E_{\text{out}} \approx 0.02$, so the correct option is **[b]**.

Problem 9

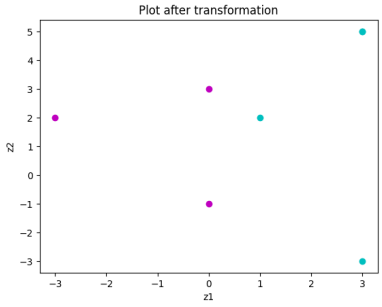
Using the code in the appendix, we get that only **[e]** is correct, E_{out} is lower using the transform than not but the improvement is less than 1%.

Problem 10

Using the code in the appendix, we get that only **[a]** is correct.

Problem 11

Using the code in the appendix, we get the plot of transformed points as:



This means that the margin maximizing plane must cut at $z_1 = 0.5$. Since z_2 does not change the label of the point, $w_2 = 0$. If the plane is given by $\langle w_1, w_2 \rangle^T z + b = 0$, on putting $z_1 = 0.5$ and $z_2 = 0$, and $w_2 = 0$, we get that $b = -0.5$. Note that we could just as easily pick any other point on the line of intersection but this was the most convenient. So the final solution is $1, 0, -0.5$ or option **[c]**.

Problem 12

Using the code in the appendix, the optimal occurs when $C \geq 10$, and the number of support vectors in this case is 5, so range 4-5, or **[c]** is correct.

Problem 13

From code in the appendix, the dataset was inseparable by the kernel RBF 0.0% of the times, so **[a]** is correct.

Problem 14

From the code in the appendix, we get that RBF kernel beats RBF regular 93% of the times, in terms of E_{out}. This means that **[e]** is the correct option.

Problem 15

From the code in the appendix, we get that for $K = 12$, RBF kernel beats RBF regular 82% of the times, which is between 60% and 90% so **[d]** is the correct option.

Problem 16

From the code in the appendix, we get that E_{in} goes down as K (`num_clusters`) goes up 83% of the times as well as E_{out} goes down as K goes up 76% of the times. Since it's most common for both to go down, **[d]** is correct.

Problem 17

From the code in the appendix, we get that E_{in} goes up as γ (`num_clusters`) goes up 52% of the times as well as E_{out} goes up as γ goes up 58% of the times. Since it's most common for both to go up, **[c]** is correct.

Problem 18

From the code in the appendix, we get that exactly 0.0% of the times is $E_{\text{in}} = 0$ achieved, so **[a]** is correct.

Problem 19

Given a single person $\mathcal{P}$ that had a heart attack, we know that $f \neq 0$. However we are not guaranteed that $f = 1$ either. This eliminates **[a]** and **[d]**. This also implies that, $P(f = 0 \mid \mathcal{P}) = 0$ and $P(0 < f \leq 1 \mid \mathcal{P}) \neq 0$. We can now determine whether $P(h = f \mid \mathcal{D})$ grows linearly or nonlinearly over $[0, 1]$. From Bayes

$$P(h = f \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid h = f)P(h = f)}{P(\mathcal{D})} \propto P(\mathcal{D} \mid h = f)P(h = f) \propto fP(h = f)$$

We can draw the final conclusion because it is obvious that for a person, the probability that they have a heart attack *given* that $h = f$ is f itself. This means that it is a linear relation, so **[b]** is correct.

Problem 20

[a] If g_1 is much closer to the actual target than g_2 then the average becomes more erroneous because of g_2-introduced error, so this is false.

[b] If If g_1 is much closer to the actual target than g_2 then the average hypothesis becomes more erroneous because of g_2-introduced error, so the average error in this case will be higher than $E_{\text{out}}(g_1)$.

[c] Could be, because of some mean squared stuff.

[d] This option is essentially the same as **[b]**, which we eliminated.

So finally, the answer is either option **[c]** or **[e]**.

Appendix

[125]:
```
import numpy as np
import math
from sklearn import svm, model_selection
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.linear_model import Ridge
from sklearn.linear_model import RidgeClassifier
from sklearn.kernel_ridge import KernelRidge
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import random
```

[46]:
```
train_data = np.loadtxt('features_final.train')
test_data = np.loadtxt('features_final.test')
```

[47]:
```
x_train = train_data[:,1:]
y_train = train_data[:,0]

x_test = test_data[:,1:]
y_test = test_data[:,0]
```

[48]:
```
def transform(xvals, toTransform=False):
    new_x = []
    if not toTransform:
        for row in xvals:
            x1 = row[0]
            x2 = row[1]
            new_x.append(np.asarray([1, x1, x2]))
    else:
        for row in xvals:
            x1 = row[0]
            x2 = row[1]
            new_x.append(np.asarray([1, x1, x2, x1*x2, x1**2, x2**2]))
    return np.asarray(new_x)
```

[49]:
```
def binarize(posDigit, yvals):
    return np.array([1 if classval == posDigit else -1 for classval in yvals])
```

[81]:
```
for classval in range(10):
    x_train_bias = transform(x_train)
    x_train_transform = transform(x_train, True)
    x_test_bias = transform(x_test)
    x_test_transform = transform(x_test, True)
    binary_y_train = binarize(classval, y_train)
    binary_y_test = binarize(classval, y_test)
    model = RidgeClassifier(alpha = 1)
```

```
    print('-'*25, end = '\n')
    print('Classifier: ' + str(classval) + ' versus all')
    if classval in range(5):
        model.fit(x_train_transform, binary_y_train)
        print('Eout_trans: ' + str(1 - model.score(x_test_transform,␣
↪binary_y_test)))
    else:
        model.fit(x_train_bias, binary_y_train)
        print('Ein_bias: ' + str(1 - model.score(x_train_bias, binary_y_train)))
    print()
```

```
-------------------------
Classifier: 0 versus all
Eout_trans: 0.10662680617837572

-------------------------
Classifier: 1 versus all
Eout_trans: 0.0219232685600399

-------------------------
Classifier: 2 versus all
Eout_trans: 0.09865470852017932

-------------------------
Classifier: 3 versus all
Eout_trans: 0.08271051320378675

-------------------------
Classifier: 4 versus all
Eout_trans: 0.09965122072745392

-------------------------
Classifier: 5 versus all
Ein_bias: 0.07625840076807022

-------------------------
Classifier: 6 versus all
Ein_bias: 0.09107118365107669

-------------------------
Classifier: 7 versus all
Ein_bias: 0.08846523110684401

-------------------------
Classifier: 8 versus all
Ein_bias: 0.074338225209162
```

```
-------------------------
Classifier: 9 versus all
Ein_bias: 0.08832807570977919
```

[88]:
```
always_overfit = True
always_improves = True
no_diff = True
always_worsens = True
fiveall = 0
for classval in range(9):
    x_train_bias = transform(x_train)
    x_train_transform = transform(x_train, True)
    x_test_bias = transform(x_test)
    x_test_transform = transform(x_test, True)
    binary_y_train = binarize(classval, y_train)
    binary_y_test = binarize(classval, y_test)

    model = RidgeClassifier(alpha = 1)
    model.fit(x_train_bias, binary_y_train)
    e_in_bias = 1 - model.score(x_train_bias, binary_y_train)
    e_out_bias = 1 - model.score(x_test_bias, binary_y_test)
    model.fit(x_train_transform, binary_y_train)
    e_in_trans = 1 - model.score(x_train_transform, binary_y_train)
    e_out_trans = 1 - model.score(x_test_transform, binary_y_test)

    if e_in_trans > e_out_trans:
        always_overfit = False
    if not(e_out_trans <= 0.95*e_out_bias):
        always_improves = False
    if e_out_bias != e_out_trans:
        no_diff = False
    if not(e_out_trans >= 1.05*e_out_bias):
        always_worsens = False
    if classval == 5:
        fiveall = 100*(e_out_trans/e_out_bias)
print('Always Overfits: ', always_overfit)
print('Always Improves by 5%+: ', always_improves)
print('Never affects change on out sample error: ', no_diff)
print('Always worsens by 5%+: ', always_worsens)
print('For 5 versus all, e_out_transform is ' + str(fiveall) + '% of e_out_bias')
```

```
Always Overfits:  False
Always Improves by 5%+:  False
Never affects change on out sample error:  False
Always worsens by 5%+:  False
For 5 versus all, e_out_transform is 99.3750000000001% of e_out_bias
```

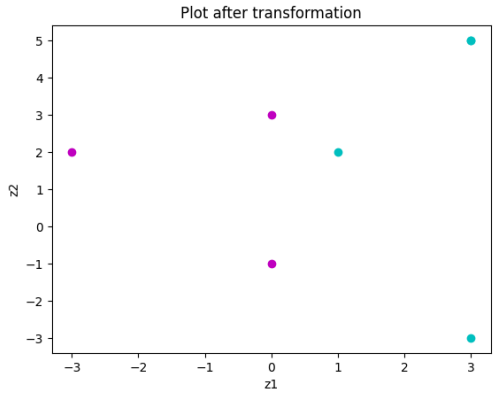
[74]:
```
x_train_new = np.asarray([x_train[i] for i in range(len(x_train)) if y_train[i]␣
↪in [1,5]])
y_train_p = [y_train[i] for i in range(len(y_train)) if y_train[i] in[1,5]]
y_train_new = np.asarray([y_train_p[i] if y_train_p[i] == 1 else -1 for i in␣
↪range(len(y_train_p))])

x_test_new = np.asarray([x_test[i] for i in range(len(x_test)) if y_test[i] in␣
↪[1,5]])
y_test_p = [y_test[i] for i in range(len(y_test)) if y_test[i] in[1,5]]
y_test_new = np.asarray([y_test_p[i] if y_test_p[i] == 1 else -1 for i in␣
↪range(len(y_test_p))])

x_train_one = transform(x_train_new)
x_train_two = transform(x_train_new, True)
x_test_one = transform(x_test_new)
x_test_two = transform(x_test_new, True)
```

[78]:
```
lambdav = [1, 0.01]
for l in lambdav:
    print('-'*25)
    print('At lambda = ' + str(l))
    model = RidgeClassifier(alpha = l)
    model.fit(x_train_two, y_train_new)
    print('Ein: ' + str(1 - model.score(x_train_two, y_train_new)))
    print('Eout: ' + str(1 - model.score(x_test_two, y_test_new)))
```

```
-------------------------
At lambda = 1
Ein: 0.005124919923126248
Eout: 0.02594339622641506
-------------------------
At lambda = 0.01
Ein: 0.004484304932735439
Eout: 0.028301886792452824
```

[99]:
```
data = [[(1, 0), -1], [(0, 1), -1], [(0, -1), -1],
     [(-1, 0), 1], [(0, 2), 1], [(0, -2), 1],
     [(-2, 0), 1]]

minus_one = []
plus_one = []

for point in data:
    x1, x2 = point[0]
    z1, z2 = [x2**2 - 2*x1 - 1, x1**2 - 2*x2 + 1]
    if point[1] == -1:
        minus_one.append([z1, z2])
```

```
    else:
        plus_one.append([z1, z2])

minus_one = np.asarray(minus_one)
plus_one = np.asarray(plus_one)

plt.plot(plus_one[:, 0], plus_one[:, 1], 'co')
plt.plot(minus_one[:, 0], minus_one[:, 1], 'mo')
plt.xlabel('z1')
plt.ylabel('z2')
plt.title("Plot after transformation")
plt.show()
```



[107]:
```
xvals = []
yvals = []
for point in data:
    xvals.append(point[0])
    yvals.append(point[1])
for exp in range(-5,10):
    model = svm.SVC(kernel = 'poly', degree = 2, C=10**exp)
    model.fit(xvals, yvals)
    print('Num support vectors at C = 10e' + str(exp) +': ' + str(sum(model.
 ↪n_support_)))
```

```
Num support vectors at C = 10e-5: 6
Num support vectors at C = 10e-4: 6
Num support vectors at C = 10e-3: 7
Num support vectors at C = 10e-2: 7
Num support vectors at C = 10e-1: 7
Num support vectors at C = 10e0: 7
Num support vectors at C = 10e1: 5
Num support vectors at C = 10e2: 5
Num support vectors at C = 10e3: 5
```

```
Num support vectors at C = 10e4: 5
Num support vectors at C = 10e5: 5
Num support vectors at C = 10e6: 5
Num support vectors at C = 10e7: 5
Num support vectors at C = 10e8: 5
Num support vectors at C = 10e9: 5
```

[115]:
```
def f(x):
    x_1, x_2 = x
    val = x_2 - x_1 + 0.25*math.sin(math.pi * x_1)
    if val == abs(val):
        return +1
    else:
        return -1
```

[152]:
```
def generate_data(N):
    dataset_x = []
    dataset_y = []
    for i in range(N):
        x_1 = random.uniform(-1, 1)
        x_2 = random.uniform(-1, 1)
        y = f([x_1, x_2])
        dataset_x.append([x_1, x_2])
        dataset_y.append(y)
    return [dataset_x, dataset_y]
```

[117]:
```
num_experiments = 1000
num_inseparable = 0
for i in range(num_experiments):
    x, y = generate_data(100)
    model = svm.SVC(kernel = 'rbf', C = 100000, gamma = 1.5)
    model.fit(x, y)
    if (1 - model.score(x, y)) != 0:
        num_inseparable += 1
print('Dataset was inseparable ' + str(100*num_inseparable/num_experiments) + '%␣
 ↪of the times')
```

```
Dataset was inseparable 0.0% of the times
```

[144]:
```
def RBF_REG(x,mu, gamma=1.5):
    return np.exp(-gamma*np.sum((x-mu)**2))

def RBF_N(xs, ct, gamma):
    z = np.zeros((xs.shape[0], 1+ct.shape[0]))
    z[:,0] = np.ones(xs.shape[0])
    for i in range(ct.shape[0]):
        z[:,i+1] = np.apply_along_axis(RBF_REG, 1, xs, ct[i], gamma)
```

```
    return z

def get_weight(P, PT, l, y):
    inter = np.matmul(np.linalg.inv(np.matmul(PT, P)
                    + l*np.identity(len(np.matmul(PT, P)))), PT)
    w_final = np.matmul(inter, y)
    return w_final

def regRBF_N(xs, yn, n_clusters, gamma):
    km = KMeans(n_clusters=n_clusters).fit(xs)
    ct = km.cluster_centers_
    z = RBF_N(xs, ct, gamma=gamma)
    w = get_weight(z, z.T, 0, yn)
    y_pred = np.dot(z, w)
    Ein = np.sum(y_pred*yn < 0)/(1.0*yn.size)
    return Ein, w, ct
```

[179]:
```
def experiment(num_clusters, gamma=1.5):
    d1 = generate_data(200)
    d2 = generate_data(200)
    xn, yn = np.asarray(d1[0]), np.asarray(d1[1])
    xn_test, yn_test = np.asarray(d2[0]), np.asarray(d2[1])
    Ein_reg, w, centers = regRBF_N(xn, yn, num_clusters, gamma=gamma)
    z_out = RBF_N(xn_test, centers, gamma=gamma)
    y_pred = np.dot(z_out, w)
    Eout_reg = np.sum(y_pred*yn_test < 0)/(200.0)
    model = svm.SVC(kernel = 'rbf', C = 1000000)
    model.fit(xn, yn)
    y_pred_kern = model.predict(xn_test)
    Ein_kern = np.sum(y_pred_kern*yn<0)/(200.0)
    Eout_kern = np.sum(y_pred_kern*yn_test<0)/(200.0)
    return Ein_reg, Eout_reg, Ein_kern, Eout_kern
```

[180]:
```
count = 0
for i in range(100):
    a, Eout_reg, b, Eout_kern = experiment(num_clusters=9)
    if Eout_reg  > Eout_kern:
        count += 1

print('Kernel beats regular RBF ' + str(count) + '% of the times')
```

```
Kernel beats regular RBF 93% of the times
```

[181]:
```
count = 0
for i in range(100):
    a, Eout_reg, b, Eout_kern = experiment(num_clusters=12)
    if Eout_reg  > Eout_kern:
        count += 1
```

```
print('Kernel beats regular RBF ' + str(count) + '% of the times')
```

```
Kernel beats regular RBF 82% of the times
```

[183]:
```
e_intruth = []
e_outruth = []
for i in range(100):
    ein9, eout9, _,_ = experiment(9)
    ein12, eout12, _,_ = experiment(12)
    e_intruth.append(ein9 < ein12)
    e_outruth.append(eout9 < eout12)
print('Ein goes up with num_clusters ' + str(e_intruth.count(True)) +'% times')
print('Ein goes down with num_clusters ' + str(e_intruth.count(False)) +'%␣
 ↪times')
print('Eout goes up with num_clusters ' + str(e_outruth.count(True)) +'% times')
print('Eout goes down with num_clusters ' + str(e_outruth.count(False)) +'%␣
 ↪times')
```

```
Ein goes up with num_clusters 17% times
Ein goes down with num_clusters 83% times
Eout goes up with num_clusters 24% times
Eout goes down with num_clusters 76% times
```

[184]:
```
e_intruth = []
e_outruth = []
for i in range(100):
    ein9, eout9, _,_ = experiment(9, 1.5)
    ein12, eout12, _,_ = experiment(9, 2)
    e_intruth.append(ein9 < ein12)
    e_outruth.append(eout9 < eout12)
print('Ein goes up with gamma ' + str(e_intruth.count(True)) +'% times')
print('Ein goes down with gamma ' + str(e_intruth.count(False)) +'% times')
print('Eout goes up with gamma ' + str(e_outruth.count(True)) +'% times')
print('Eout goes down with gamma ' + str(e_outruth.count(False)) +'% times')
```

```
Ein goes up with gamma 52% times
Ein goes down with gamma 48% times
Eout goes up with gamma 58% times
Eout goes down with gamma 42% times
```

[188]:
```
ct = 0
for i in range(100):
    ein9, eout9, _,_ = experiment(9, 1.5)
    if ein9 ==0:
        ct += 1
print('Ein is 0 a total ' + str(ct) +'% times')
```

```
Ein is 0 a total 0% times
```