Warm-Up:

1.

This is to prove:

for all x, y ∈ Rn and λ∈[0, 1], λ max{f1(x), f2(x)} + (1 - λ) max{f1(y), f2(y)} >= max{f1(λx + (1 - λ)y), f2(λx + (1 - λ)y)}

∵f1 and f1 are convex functions

∴λf1(x) + (1 - λ) f1(y) >= f1(λx + (1 - λ)y) and λf2(x) + (1 - λ) f2(y) >= f2(λx + (1 - λ)y)

Use following two inequality:

1. max(a, b) >= a
2. When a >= b, a >= c then a >= max(b, c)

Using inequality 1:

λ max{f1(x), f2(x)} + (1 - λ) max{f1(y), f2(y)} >= λf1(x) + (1 - λ) f1(y) >= f1(λx + (1 - λ)y)

λ max{f1(x), f2(x)} + (1 - λ) max{f1(y), f2(y)} >= λf2(x) + (1 - λ) f2(y) >= f2(λx + (1 - λ)y)

Using inequality 2:

λ max{f1(x), f2(x)} + (1 - λ) max{f1(y), f2(y)} >= max{f1(λx + (1 - λ)y), f2(λx + (1 - λ)y)}

This is what we want to prove.

2.a

f(x) = max{x2 - 2x, |x|} at x = 0 and x = -2

For x = 0, when x = 0, x2 - 2x = 0 = |x|

f(x) = x2 - 2x when x<=0, |x| when 0<=x<=3

So the function at 0 is not derivable

When x-🡪0 f’(x) = 2x – 2, f’(0) = -2

When x+🡪0 f’(x) = 1, f’(0) = 1

So the set of subgradient’s differential at 0 is [-2,1]

One of subgradient is y = 0

For x = -2, when x = -2, x2 - 2x = 8 != 2 = |x|

f(x) = x2 - 2x when x<=0

So the Derivative at -2 is derivable

When x🡪-2 f’(x) = 2x – 2, f’(-2) = -6

So the value of subgradient’s differential at -2 is -6

The subgradient is y = -6x - 4

2.b

g(x) = max{(x - 1)2, (x - 2)2} at x = 1.5 and x = 0.

For x = 1.5, when x = 1.5, (x - 1)2 = 0.25 = (x - 2)2

f(x) = (x - 2)2 when x<= 1.5, f(x) = (x - 1)2 when x>=1.5

So the function at 1.5 is not derivable

When x-🡪1.5 f’(x) = 2x – 4, f’(1.5) = -1

When x+🡪0 f’(x) = 2x - 2, f’(1.5) = 1

So the set of subgradient’s differential at 0 is [-1,1]

One of subgradient is y = 0.25

For x = 0, when x = 0, (x - 1)2 = 1 != 4 = (x - 2)2

f(x) = (x - 2)2 when x<= 1.5

So the function at 0 is derivable

When x🡪0 f’(x) = 2x – 4, f’(0) = -4

So the value of subgradient’s differential at 0 is -4

The subgradient is y = -4x + 4

Problem 1:

1.

The number of iterations:

46

First three steps’ w and b:

[1278.996461083001, 460.0612580119997, -108.55851403599995, -1672.3157294790021]T -354.0

[1307.294729741001, 432.7477879899997, -27.55191988299994, -1523.7889544580023]T -493.0

[1255.189813620001, 425.5040288239997, 18.79654039900003, -1434.6675419700023]T -625.0

Final weights:

[685.7993289150007, 243.89947473099983, 8.241991933000026, -797.6250531350023] -1485.0

Bias:

0.0

2.

The number of iterations:

8694

First three steps’ w and b:

[4.617544237, 2.469679381, 1.967660787, -1.813355506] T -1.0 [3.4532228789999997, 0.16943481800000004, 2.62801595, -4.647098506] T -2.0

[0.45610583999999976, 4.929004722, -2.2588903150000004, -4.651540471] T -3.0

Final weights:

[149.27714019000345, 52.53347316700037, 1.6716726529996597, -172.89194014099422] T -322.0

Bias:

0.0

3.

For Standard gradient descent,

when step size γt = 100.0, the number of iterations that it takes to find a perfect classifier for the data is 46.

when step size γt = 1.0, the number of iterations that it takes to find a perfect classifier for the data is 46.

when step size γt = 0.01, the number of iterations that it takes to find a perfect classifier for the data is 46.

So the change the step size has no effect on rate of convergence change.

Because for classifier equation 𝑤𝑇𝑥 + 𝑏 = 0, when you multiply any γt with wT and b at each steps, it make no difference for the equation.

4.

The smallest number of data points in a two-dimensional data set fail to converge is 3. The 3 points in same line segment, the middle one’s class label different from the other two.

In two-dimensional data set, when there is not a line can separate the two class labels, the algorithm fails to converge. Generally, when there is not a function 𝑤𝑇𝑥 + 𝑏 = 0 can separate the labels in a dataset, the algorithm will fail to converge. But when we mean to converge to specific level(abs(grada) \* abs(gradb) < epsilon), the algorithm can still work for more general cases. It also may fail for more discrete dataset.

Problem 2:

1.a

Φ(x1,x2) = [x1 + x2, x1 – x2]

not separable.

Justify:

Label + data set is (-2,0) (2,0)

Label – data set is (0,-2) (0,2)

The line segment between label + data set join with the line segment between label - data set, so there is no linear separator can separate them.

1.b

Φ(x1,x2) = [exp(x1), exp(x2)]

not separable.

Justify:

Label + data set is (1/e, 1/e) (e, e)

Label – data set is (1/e, e) (e, 1/e)

The line segment between label + data set join with the line segment between label - data set, so there is no linear separator can separate them.

1.c

Φ(x1,x2) = [x12, x22, x1 x2]

Separable

Justify:

Label + data set is (1, 1, 1)

Label – data set is (1, 1, -1)

The function separate these two points is [0,0,1]T *x* = 0

1.d

Φ(x1,x2) = [x1 sin(x2), x1]

Separable

Justify:

Label + data set is (sin1, -1), (sin1, 1)

Label – data set is (-sin1, -1), (-sin1, 1)

The function separate these two points is [1,0]T *x* = 0

2.

[𝑤1,w2,…,wk]𝑇[x1, x2, x12, x22, x1x2,…,x1mx2n]+ 𝑏 = 0

There are totally k elements in [x1, x2, x12, x22, x1x2,…,x1cx2d], c>=0, d >= 0, and k<= <= k + c +d

Let The size of feature representation be k, let the number of training data points be n, then

The per iteration complexity of standard gradient descent is O(kn).

Problem 3:

1.

The Piecewise linear function a1x+ b1 and a2x+ b2 should join in specific point(x0,y0), if they parallel , topic return to linear regression. The x0 can be calculated from a1x+ b1 = a2x+ b2, x0 = (b2 – b1)/(a1 – a2), so the minimum loss function is:

L(a1,b1,a2,b2) = min(a1,b1,a2,b2) 1/𝑀 (max{a1xi+ b1, a2xi+ b2} – yi)2

= 1/𝑀 ( (min(a1,b1) (a1xi + b1 – yi)2 | xi <= x0) +

(min(a2,b2) (a2xi+ b2 – yi)2 | xi >= x0) )

2.

Let γ(t) be the t step’s step size for linear regression.

L’(a1) = 1/𝑀2(a1xi + b1 – yi) xi | xi <= x0)

L’(b1) = 1/𝑀2(a1xi + b1 – yi)| xi <= x0)

L’(a2) = 1/𝑀2(a2xi + b2 – yi) xi | xi >= x0)

L’(b2) = 1/𝑀2(a2xi + b2 – yi)| xi >= x0)

a1(t+1) = a1(t) - γ(t) (1/𝑀2(a1xi + b1 – yi) xi | xi <= x0))

b1(t+1) = b1(t) - γ(t) (1/𝑀2(a1xi + b1 – yi)| xi <= x0))

a2(t+1) = a2(t) - γ(t) (1/𝑀2(a2xi + b2 – yi) xi | xi >= x0))

b2(t+1) = b2(t) - γ(t) (1/𝑀2(a2xi + b2 – yi)| xi >= x0))

Set any random a1 b1 a2 b2 at begin, then use formula above to calculate a1 b1 a2 b2 iteratively. When a1 b1 a2 b2 stop changing (or close to stop changing in some cases), we get result a1 b1 a2 b2.

3.

It is convex.

When a12xi2 > 0, b12 > 0 and a22xi2 > 0, b22 > 0, both of the two Quadratic polynomial f(a1,b1) = (a1xi + b1 – yi)2 and f(a2,b2) = (a2xi + b2 – yi)2 should be convex function.

Use the conclusion from Warm-Up question1, two convex functions’ maximum function is also convex function.

So the optimization problem function I used here is convex function.

Problem 3:

I use python library sklearn.svm to solve this problem (actually I hoped to realize my own SVM, but cannot finish before deadline). It’s a perfect classifier that can reach accuracy 1.0.

The parameters can be seen from follows:

<bound method BaseEstimator.get\_params of SVC(C=30, cache\_size=200, class\_weight=None, coef0=0.0,decision\_function\_shape='ovr', degree=3, gamma=20, kernel='rbf',max\_iter=-1, probability=False, random\_state=None, shrinking=True,tol=0.001, verbose=True)>

I only optimize C for Penalty parameter, gamma for Kernel coefficient (higher value of each parameters will reach high accuracy, it may also lead to over-fitting).

When C=24, gamma=20 the accuracy: 1.0

When C=24, gamma=19 the accuracy: 0.999

When C=23, gamma=20 the accuracy: 0.999

There is not any function in sklearn library that give us optimal margin accurately, so I just get a round boundary with precision 1.0. From the data shown above, there is a boundary when we reach perfect classifier, it is in 23 <= C <= 24 and 19 <= gamma <= 20.

The line of support vectors are listed as follows(start from 0):

[ 0 7 23 61 99 121 130 144 152 154 155 156 163 173 178 185 193 200

206 213 220 240 242 249 261 276 279 321 344 345 389 428 474 503 507 509

513 530 553 571 594 636 677 681 688 692 707 708 710 717 719 769 771 773

774 787 796 799 801 802 824 849 855 857 867 900 907 913 918 931 940 946

947 956 972 985 3 4 9 10 13 16 19 22 24 33 35 38 40 41

53 55 65 67 70 71 77 78 82 87 105 106 110 112 114 115 117 118

123 125 127 129 132 139 159 161 165 166 169 170 172 188 189 191 195 196

198 205 207 214 215 216 217 219 221 223 224 225 227 233 235 236 243 250

252 265 270 288 297 299 300 304 305 306 309 312 316 318 320 324 326 329

332 337 338 339 341 348 349 350 351 354 358 363 367 369 376 386 388 394

401 407 413 415 426 427 433 435 442 444 450 454 456 457 463 466 478 484

485 492 502 505 508 512 516 519 526 529 532 536 539 540 547 550 551 558

562 563 564 565 567 573 578 582 585 589 592 608 609 613 614 615 616 617

619 633 644 646 647 654 658 661 662 669 672 676 678 679 683 684 689 713

721 726 744 745 748 761 764 766 777 781 784 785 790 793 795 798 813 819

825 834 835 836 839 842 844 847 850 854 858 859 864 865 868 873 874 876

879 884 888 891 895 898 901 903 904 911 912 914 937 943 944 945 950 955

957 970 978 982 984 986 989 992 993 994 996]