Deadline: Friday. December 4th , 10:00 Please upload a `.pdf` file via LearnWeb. (e.g. exported Jupyter notebook)

1 L^2 regularization (10 points)

A [7p] Fit a linear regression model (including bias) with L^2 regularization to the dataset from Table **??** by performing 2 iterations of coordinate descent (update each parameter twice). Use $\beta^{(0)} = 0$ and $\lambda = 0.5$.

x_1	x_2	y
1	1	1.4
1	-1	1.6
-1	0	0.5
-1	-1	0.6

Table 1

B [3p] The **elastic-net** model is a linear model with a mix of L^1 and L^2 regularization.

$$L^{\text{enet}}(\beta) = \frac{1}{2N}\|y - X\beta\|_2^2 + \lambda\Big(\alpha\|\beta\|_1 + (1-\alpha)\tfrac{1}{2}\|\beta\|_2^2\Big)$$

Note that if $\alpha = 1$, elastic net is the same as LASSO and for $\alpha = 0$ it is the same as RIDGE regression. For $\alpha \in (0, 1)$ it is something in between. We trained an Elastic Net model 4 times on a regression task, each time choosing a different trade-off $\alpha \in \{0, 0.25, 0.5, 1\}$. The resulting regularization paths, as well as the number of non-zero coefficients at different total regularization strength λ is shown in Figure **??**. Explain which figure corresponds to which choice of α.

material/lassopaths.pdf

Figure 1: Regularization paths of the 4 models

2 Hyperparameter Optimization – Programming (10 points)

Use the following code to load the ”IRIS” dataset using the sklearn library. Follow the TODOs.

```
from sklearn.svm import SVC
from sklearn import datasets
from sklearn.modelselection import traintestsplit
from sklearn.metrics import SCORERS
from sklearn.modelselection import crossvalscore
from sklearn.metrics import accuracyscore

CVSPLITS=5

data, target = datasets.loadiris(returnXy=True)

shuffleseed = 2020
# Always shuffle your data to be safe. Use fixed seed for reprod.
```

```
dataX, dataXt, datay, datayt = traintestsplit(
    data, target, testsize=0.2, randomstate=shuffleseed, shuffle=True
)

hyperparameters = -
    ”C”: -
        ”range”: (1.0, 1e3)
        ,
    ”gamma”: -
        ”range”: (1e-4, 1e-3)
        ,


fixedparameters =  -
    ”kernel”: ”rbf”,
    ”probability”: True,
    ”tol”:1e-1


# TODO : Select 100 pairs of hyperparameters, e.g. -”C”:4,”gamma”:2e-4
# Iteratively:
#           TODO : create a parameters dictionary including the fixedparameters
     and the new hyper-parameters
#           TODO : Define a SVC Model given the new parameters
#           clf = ?
#           Do a cross validation and report the mean and standard deviation
#           S = crossvalscore(clf, dataX, datay, scoring=SCORERS[”accuracy
    ”], cv=CVSPLITS)
#           Report the test accuracy
# Visualize the results on a 2D grid. Show one figure for the validation, and
   one figure for the test results.
```

[5]Parameter Variance – OLS vs Ridge Regression For the following problem, we assume that the ground truth is is a linear function $y(x) = x^\intercal\hat{\beta} + \epsilon$ with $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and we are given a **finite** data sample (X, Y). From the lecture we know that the ordinary least squares (OLS) estimator $\hat{\beta}^{\text{OLS}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ satisfies:

- $\mathbb{E}[\hat{\beta}^{\text{OLS}}] = \hat{\beta}$
- $\mathbb{V}[\hat{\beta}^{\text{OLS}}] = (X^\mathsf{T}X)^{-1}\sigma^2$

In particular, we note that the OLS estimator is unbiased!

A [2p] Show that the RIDGE estimator $\hat{\beta}^{\text{RIDGE}} = (X^\mathsf{T}X + \lambda\,\mathbb{I})^{-1}X^\mathsf{T}y$ satisfies

- $\mathbb{E}[\hat{\beta}^{\text{RIDGE}}] = (X^\mathsf{T}X + \lambda\,\mathbb{I})^{-1}X^\mathsf{T}X\hat{\beta}$
- $\mathbb{V}[\hat{\beta}^{\text{RIDGE}}] = (X^\mathsf{T}X + \lambda\,\mathbb{I})^{-1}X^\mathsf{T}X(X^\mathsf{T}X + \lambda\,\mathbb{I})^{-1}\sigma^2$

In particular, we note that the RIDGE estimator is biased!

B [3p] Given two covariance matrices Σ_A and Σ_B, we say that Σ_A is strictly greater than Σ_B (in symbols $\Sigma_A > \Sigma_B$) iff $\Sigma_A - \Sigma_B$ is positive definite. (This is the so called Löwner order). Show that $\hat{\beta}^{\text{OLS}}$ has stricly greater variance than $\hat{\beta}^{\text{RIDGE}}$

Hint: Note that $(X^\mathsf{T}X)^{-1}$ and $X^\mathsf{T}X + \lambda\,\mathbb{I}$ commute. More generally, if p and q are polynomial functions, then $p(A)q(A) = q(A)p(A)$ and likewise $q(A)^{-1}p(a) = p(A)q(A)^{-1}$ for any square matrix A.