Part 1: The James-Stein Estimator (20 P)

Let $x_1, \ldots, x_N \in \mathbb{R}^d$ be independent draws from a multivariate Gaussian distribution with mean vector μ and covariance matrix $\Sigma = \sigma^2 I$. It can be shown that the maximum-likelihood estimator of the mean parameter μ is the empirical mean given by:

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Maximum-likelihood appears to be a strong estimator. However, it was demonstrated that the following estimator

$$\hat{\mu}_{JS} = \left(1 - \frac{(d-2) \cdot \frac{\sigma^2}{N}}{\|\hat{\mu}_{\mathrm{ML}}\|^2}\right) \hat{\mu}_{\mathrm{ML}}$$

(a shrinked version of the maximum-likelihood estimator towards the origin) has actually a smaller distance from the true mean when $d \geq 3$. This however assumes knowledge of the variance of the distribution for which the mean is estimated. This estimator is called the James-Stein estimator. While the proof is a bit involved, this fact can be easily demonstrated empirically through simulation. This is the object of this exercise.

The code below draws ten 50-dimensional points from a normal distribution with mean vector $\mu = (1, \ldots, 1)$ and covariance $\Sigma = I$.

In [100…

```
import numpy

def getdata(seed):

    n = 10              # data points
    d = 50              # dimensionality of data
    m = numpy.ones([d]) # true mean
    s = 1.0             # true standard deviation

    rstate = numpy.random.mtrand.RandomState(seed)
    X = rstate.normal(0,1,[n,d])*s+m

    return X,m,s
```

The following function computes the maximum likelihood estimator from a sample of the data assumed to be generated by a Gaussian distribution:

In [101…
```
def ML(X):
    return X.mean(axis=0)
```

Implementing the James-Stein Estimator (10 P)

- **Based on the ML estimator function, write a function that receives as input the data $(X_i)_{i=1}^n$ and the (known) variance σ^2 of the generating distribution, and computes the James-Stein estimator**

In [102…
```
def JS(X,s):
    # REPLACE BY YOUR CODE
    u_ML = ML(X)
    N = X.shape[0]
    d = X.shape[1]
    norm_2 = (numpy.linalg.norm(u_ML))**2

    m_JS = (1 - ((d - 2)*((s**2)/N))/norm_2) * u_ML
    ###
    return m_JS
```

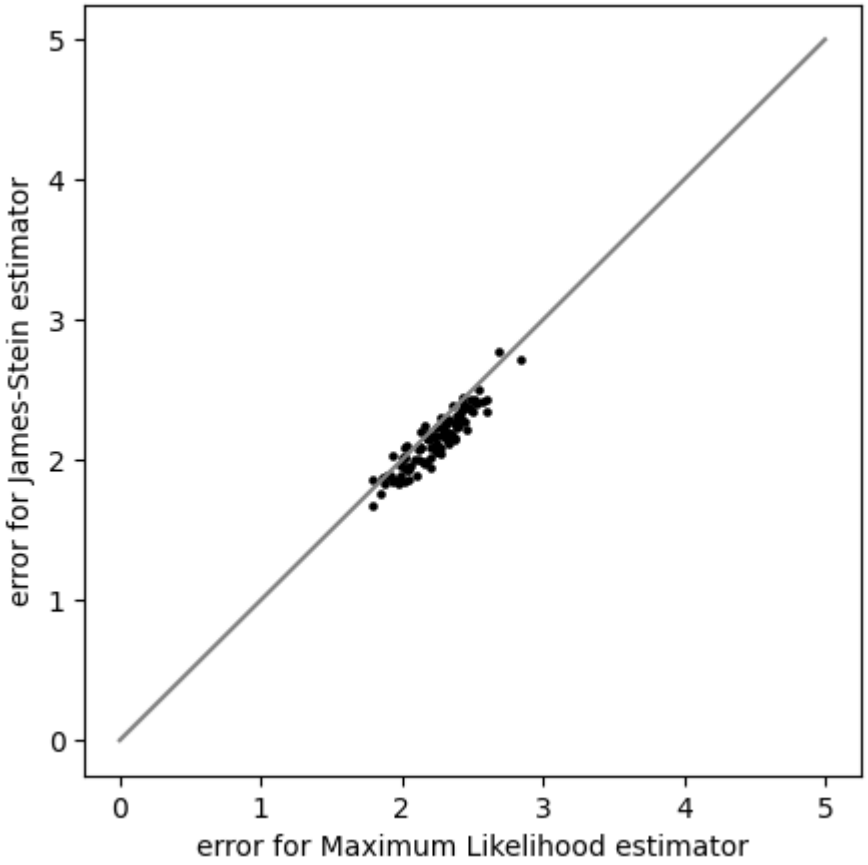
Comparing the ML and James-Stein Estimators (10 P)

We would like to compute the error of the maximum likelihood estimator and the James-Stein estimator for 100 different samples (where each sample consists of 10 draws generated by the function `getdata` with a different random seed). Here, for reproducibility, we use seeds from 0 to 99. The error should be measured as the Euclidean distance between the true mean vector and the estimated mean vector.

- **Compute the maximum-likelihood and James-Stein estimations.**
- **Measure the error of these estimations.**
- **Build a scatter plot comparing these errors for different samples.**

In [103…
```
%matplotlib inline
import matplotlib.pyplot as plt
### REPLACE BY YOUR CODE
```

```
plt.figure(figsize=(5,5))

for i in range(100):
    X, m, s = getdata(i)
    u_ML = ML(X)
    u_JS = JS(X, s)
    error_ML = numpy.linalg.norm(m - u_ML)
    error_JS = numpy.linalg.norm(m - u_JS)
    plt.scatter(error_ML, error_JS, s=5, color='black')

plt.xlabel('error for Maximum Likelihood estimator')
plt.ylabel('error for James-Stein estimator')

plt.plot([0, 5], [0, 5], color='gray')
plt.show()

###
```



Part 2: Bias/Variance Decomposition (30 P)

In this part, we would like to implement a procedure to find the bias and variance of different predictors. We consider one for regression and one for classification. These predictors are available in the module utils.

- **`utils.ParzenRegressor`:** A regression method based on Parzen window. The hyperparameter corresponds to the scale of the Parzen window. A large scale creates a more rigid model. A small scale creates a more flexible one.
- **`utils.ParzenClassifier`:** A classification method based on Parzen window. The hyperparameter corresponds to the scale of the Parzen window. A large scale creates a more rigid model. A small scale creates a more flexible one. Note that instead of returning a

single class for a given data point, it outputs a probability distribution over the set of possible classes.

Each class of predictor implements the following three methods:

- **`__init__(self,parameter):`** Create an instance of the predictor with a certain scale parameter.
- **`fit(self,X,T):`** Fit the predictor to the data (a set of data points `X` and targets `T`).
- **`predict(self,X):`** Compute the output values arbitrary inputs `X`.

To compute the bias and variance estimates, we require *multiple samples* from the training set for a single set of observation data. To acomplish this, we utilize the **`Sampler`** class provided. The sampler is initialized with the training data and passed to the method for estimating bias and variance, where its function **`sampler.sample()`** is called repeatedly in order to fit multiple models and create an ensemble of prediction for each test data point.

Regression Case (15 P)

For the regression case, Bias, Variance and Error are given by:

- $\text{Bias}(Y)^2 = (\mathbb{E}_Y[Y - T])^2$
- $\text{Var}(Y) = \mathbb{E}_Y[(Y - \mathbb{E}_Y[Y])^2]$
- $\text{Error}(Y) = \mathbb{E}_Y[(Y - T)^2]$

Task: Implement the KL-based Bias-Variance Decomposition defined above. The function should repeatedly sample training sets from the sampler (as many times as specified by the argument nbsamples), learn the predictor on them, and evaluate the variance on the out-of-sample distribution given by X and T.

In [104…

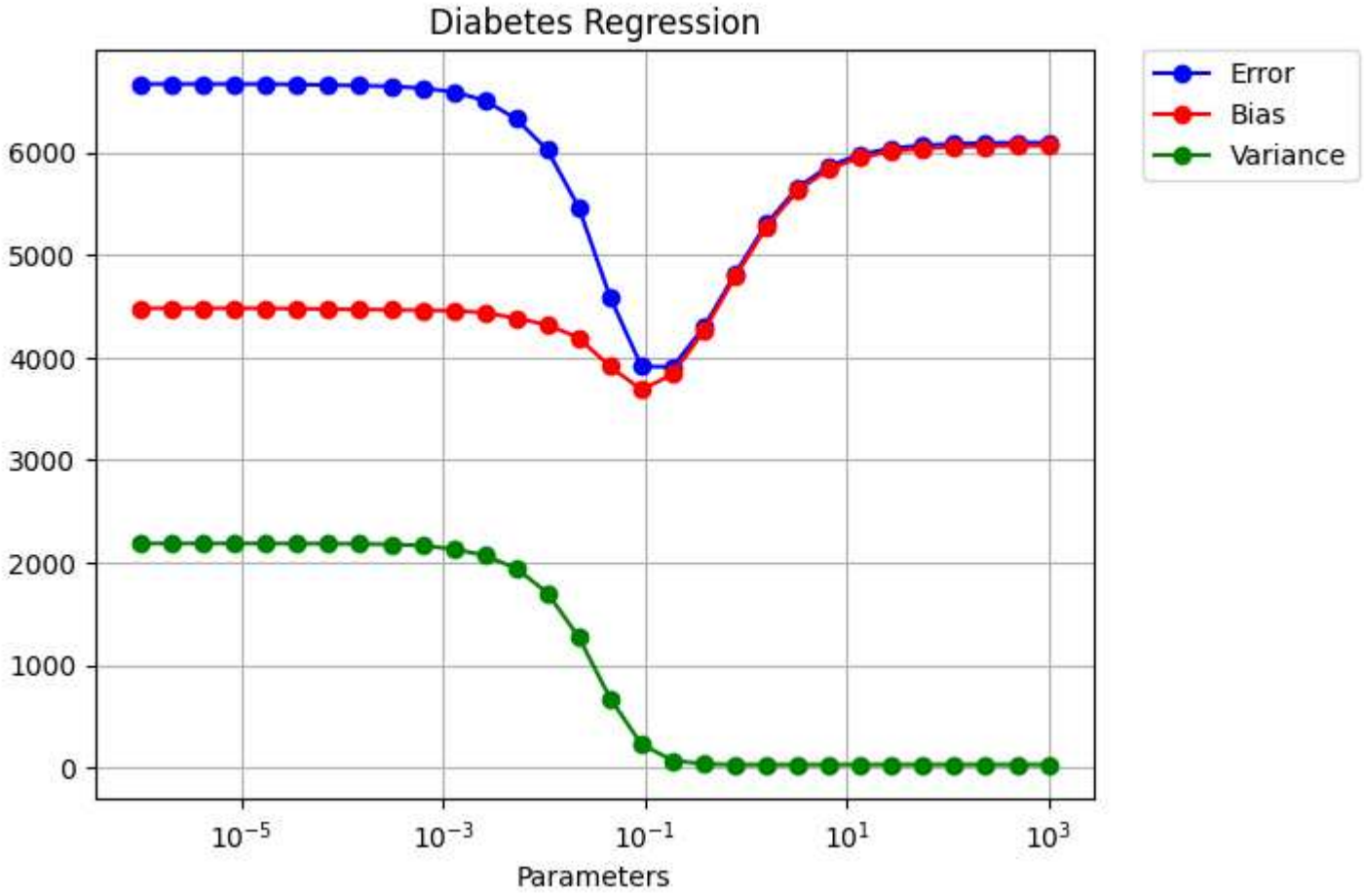
```
def biasVarianceRegression(sampler, predictor, X, T, nbsamples=25):

    # -------------------------------
    # TODO: REPLACE BY YOUR CODE
    # -------------------------------
    Y = numpy.array([predictor.fit(*sampler.sample()).predict(X) for _ in range(nbsamples)])

    Y_Mean = numpy.mean(Y, axis=0)
    bias = numpy.mean((Y_Mean - T)**2)
```

```
    # bias = numpy.mean((Y - T))**2
    variance = numpy.mean((Y - Y_Mean)**2)
    error = numpy.mean((Y - T)**2)

    # ---------------------------------

    return bias,variance
```

Your implementation can be tested with the following code:

In [105…

```
import utils,numpy
%matplotlib inline
utils.plotBVE(utils.Diabetes,numpy.logspace(-6,3,num=30),utils.ParzenRegressor,biasVarianceRegression,'Diabetes Reg
```



Classification Case (15 P)

We consider here the Kullback-Leibler divergence as a measure of classification error, as derived in the exercise, the Bias, Variance decomposition for such error is:

- $\text{Bias}(Y) = D_{\text{KL}}(T||R)$
- $\text{Var}(Y) = \mathbb{E}_Y[D_{\text{KL}}(R||Y)]$
- $\text{Error}(Y) = \mathbb{E}_Y[D_{\text{KL}}(T||Y)]$

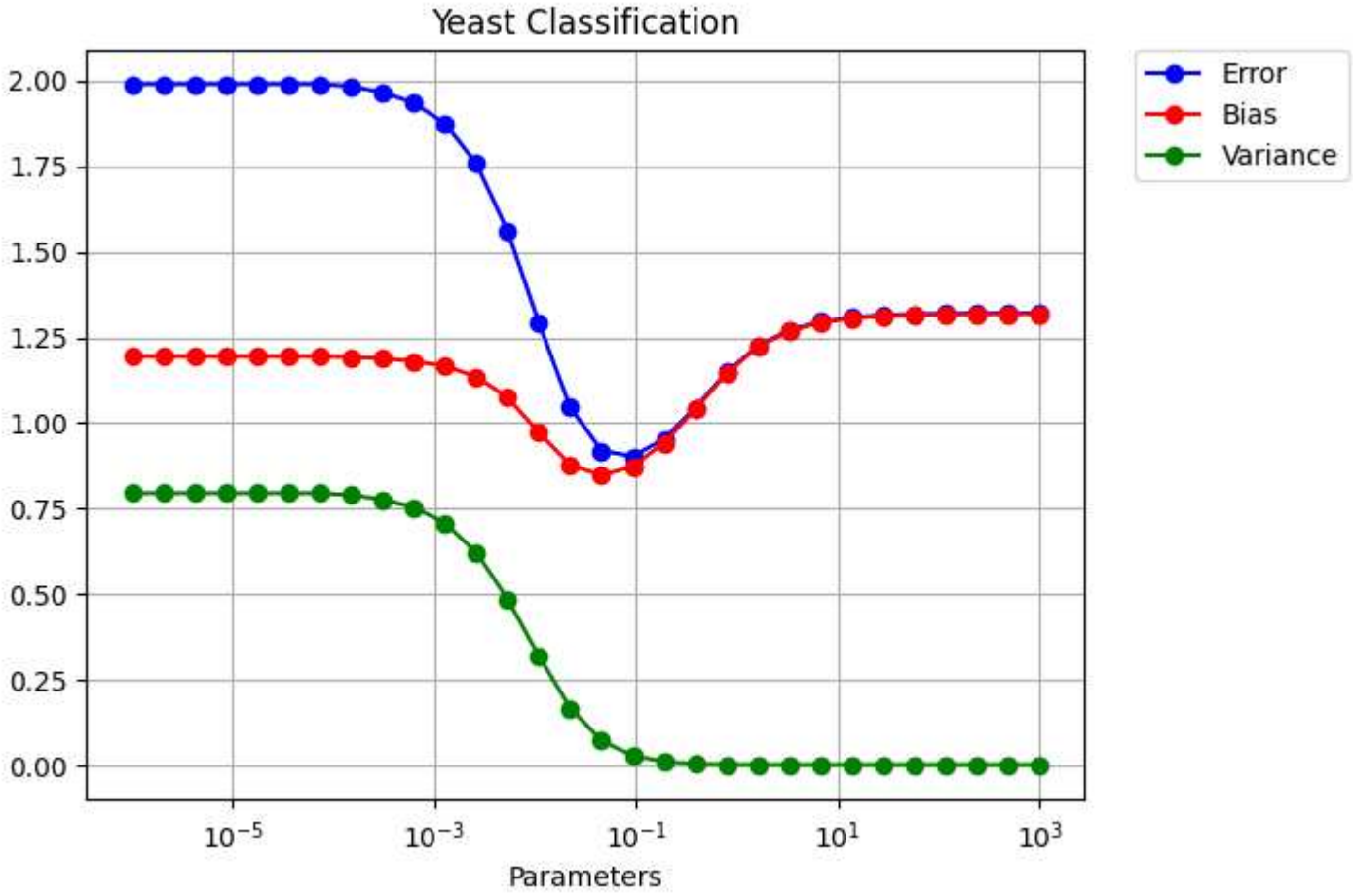
where R is the distribution that minimizes its expected KL divergence from the estimator of probability distribution Y (see the theoretical exercise for how it is computed exactly), and where T is the target class distribution.

Task: Implement the KL-based Bias-Variance Decomposition defined above. The function should repeatedly sample training sets from the sampler (as many times as specified by the argument nbsamples), learn the predictor on them, and evaluate the variance on the out-of-sample distribution given by X and T.

In [106…
```
def biasVarianceClassification(sampler, predictor, X, T, nbsamples=25):

    # ---------------------------------
    # TODO: REPLACE BY YOUR CODE
    # ---------------------------------
    hatP = numpy.array([predictor.fit(*sampler.sample()).predict(X) for _ in range(nbsamples)])
    P = T[None, :, :]

    R = numpy.exp(numpy.log(hatP).mean(axis=0, keepdims=True))
    R /= R.sum(axis=2, keepdims=True)

    bias = (P*numpy.log(P/R)).sum(axis=2).mean(axis=0).mean()
    variance = (P*numpy.log(R/hatP)).sum(axis=2).mean(axis=0).mean()
    error = (P*numpy.log(P/hatP)).sum(axis=2).mean(axis=0).mean()

    # ---------------------------------

    return bias,variance
```

Your implementation can be tested with the following code:

In [107…
```
import utils,numpy
%matplotlib inline
utils.plotBVE(utils.Yeast,numpy.logspace(-6,3,num=30),utils.ParzenClassifier,biasVarianceClassification,'Yeast Clas
```


Yeast Classification
Error
Bias
Variance
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^{-5}
10^{-3}
10^{-1}
10^{1}
10^{3}
Parameters

In [ ]:

Exercises for the course
Machine Learning 1
Winter semester 2023/24

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 7

Exercise 1: Bias and Variance of Mean Estimators (20 P)

Assume we have an estimator $\hat{\theta}$ for a parameter θ. The bias of the estimator $\hat{\theta}$ is the difference between the true value for the estimator, and its expected value

$$\text{Bias}(\hat{\theta}) = \text{E}\big[\hat{\theta} - \theta\big].$$

If $\text{Bias}(\hat{\theta}) = 0$, then $\hat{\theta}$ is called unbiased. The variance of the estimator $\hat{\theta}$ is the expected square deviation from its expected value

$$\text{Var}(\hat{\theta}) = \text{E}\big[(\hat{\theta} - \text{E}[\hat{\theta}])^2\big].$$

The mean squared error of the estimator $\hat{\theta}$ is

$$\text{Error}(\hat{\theta}) = \text{E}\big[(\hat{\theta} - \theta)^2\big] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}).$$

Let $X_1, \ldots, X_N$ be a sample of i.i.d random variables. Assume that X_i has mean μ and variance σ^2. *Calculate* the bias, variance and mean squared error of the mean estimator:

$$\hat{\mu} = \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} X_i$$

where α is a parameter between 0 and 1.

Exercise 2: Bias-Variance Decomposition for Classification (30 P)

The bias-variance decomposition usually applies to regression data. In this exercise, we would like to obtain similar decomposition for classification, in particular, when the prediction is given as a probability distribution over C classes. Let $P = [P_1, \ldots, P_C]$ be the ground truth class distribution associated to a particular input pattern. Assume a random estimator of class probabilities $\hat{P} = [\hat{P}_1, \ldots, \hat{P}_C]$ for the same input pattern. The error function is given by the expected KL-divergence between the ground truth and the estimated probability distribution:

$$\text{Error} = \text{E}\big[D_{\text{KL}}(P||\hat{P})\big] = \text{E}\big[\textstyle\sum_{i=1}^{C} P_i \log(P_i/\hat{P}_i)\big].$$

First, we would like to determine the mean of of the class distribution estimator $\hat{P}$. We define the mean as the distribution that minimizes its expected KL divergence from the the class distribution estimator, that is, the distribution R that optimizes

$$\min_R \ \text{E}\big[D_{\text{KL}}(R||\hat{P})\big].$$

(a) *Show* that the solution to the optimization problem above is given by

$$R = [R_1, \ldots, R_C] \quad \text{where} \quad R_i = \frac{\exp \text{E}\big[\log \hat{P}_i\big]}{\sum_j \exp \text{E}\big[\log \hat{P}_j\big]} \qquad \forall\ 1 \le i \le C.$$

(Hint: To implement the positivity constraint on R, you can reparameterize its components as $R_i = \exp(Z_i)$, and minimize the objective w.r.t. Z.)

(b) *Prove* the bias-variance decomposition

$$\text{Error}(\hat{P}) = \text{Bias}(\hat{P}) + \text{Var}(\hat{P})$$

where the error, bias and variance are given by

$$\text{Error}(\hat{P}) = \text{E}\big[D_{\text{KL}}(P||\hat{P})\big], \qquad \text{Bias}(\hat{P}) = D_{\text{KL}}(P||R), \qquad \text{Var}(\hat{P}) = \text{E}\big[D_{\text{KL}}(R||\hat{P})\big].$$

(Hint: as a first step, it can be useful to show that $\text{E}[\log R_i - \log \hat{P}_i]$ does not depend on the index i.)

Exercise 3: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

Exercise 1: Bias and Variance of Mean Estimators (20 P)

Assume we have an estimator $\hat{\theta}$ for a parameter θ. The bias of the estimator $\hat{\theta}$ is the difference between the true value for the estimator, and its expected value

$$\text{Bias}(\hat{\theta}) = \text{E}\big[\hat{\theta} - \theta\big].$$

If $\text{Bias}(\hat{\theta}) = 0$, then $\hat{\theta}$ is called unbiased. The variance of the estimator $\hat{\theta}$ is the expected square deviation from its expected value

$$\text{Var}(\hat{\theta}) = \text{E}\big[(\hat{\theta} - \text{E}[\hat{\theta}])^2\big].$$

The mean squared error of the estimator $\hat{\theta}$ is

$$\text{Error}(\hat{\theta}) = \text{E}\big[(\hat{\theta} - \theta)^2\big] = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta}).$$

Let $X_1, \ldots, X_N$ be a sample of i.i.d random variables. Assume that X_i has mean μ and variance σ^2. *Calculate* the bias, variance and mean squared error of the mean estimator:

$$\hat{\mu} = \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} X_i$$

where α is a parameter between 0 and 1.

Solution:

(1) Bias:

$$\text{Bias}(\hat{\mu}) = E[\hat{\mu} - \mu] = E\left[\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right]$$

$$= \alpha\, E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] - \mu$$

$$= (\alpha - 1)\mu$$

(2) Variance:

$$\text{Var}(\hat{\mu}) = E\left[(\hat{\mu} - E[\hat{\mu}])^2\right]$$

$$= \text{Var}\left(\alpha \cdot \frac{1}{N}\sum_{i=1}^{N} X_i\right)$$

$$= \frac{\alpha^2}{N^2}\,\text{Var}\left(\sum_{i=1}^{N} X_i\right)$$

Since $X_1, X_2 \cdots X_N$ are i.i.d variables

$$\therefore \text{Var}(\hat{\mu}) = \frac{\alpha^2}{N^2}\sum_{i=1}^{N}\text{Var}(X_i)$$

$$= \frac{\alpha^2}{N^2} \cdot N \cdot \sigma^2$$

$$= \frac{\alpha^2\sigma^2}{N}$$

(3) Error

$$\text{Error}(\hat{\mu}) = \text{Bias}^2(\hat{\mu}) + \text{Var}(\hat{\mu})$$

$$= (\alpha - 1)^2\mu^2 + \frac{\alpha^2}{N}\sigma^2$$

Exercise 2: Bias-Variance Decomposition for Classification (30 P)

The bias-variance decomposition usually applies to regression data. In this exercise, we would like to obtain similar decomposition for classification, in particular, when the prediction is given as a probability distribution over C classes. Let $P = [P_1, \ldots, P_C]$ be the ground truth class distribution associated to a particular input pattern. Assume a random estimator of class probabilities $\hat{P} = [\hat{P}_1, \ldots, \hat{P}_C]$ for the same input pattern. The error function is given by the expected KL-divergence between the ground truth and the estimated probability distribution:

$$\text{Error} = \mathrm{E}\big[D_{\mathrm{KL}}(P||\hat{P})\big] = \mathrm{E}\big[\textstyle\sum_{i=1}^{C} P_i \log(P_i/\hat{P}_i)\big].$$

First, we would like to determine the mean of of the class distribution estimator $\hat{P}$. We define the mean as the distribution that minimizes its expected KL divergence from the the class distribution estimator, that is, the distribution R that optimizes

$$\min_R \ \mathrm{E}\big[D_{\mathrm{KL}}(R||\hat{P})\big].$$

(a) *Show* that the solution to the optimization problem above is given by

$$R = [R_1, \ldots, R_C] \quad \text{where} \quad R_i = \frac{\exp \mathrm{E}\big[\log \hat{P}_i\big]}{\sum_j \exp \mathrm{E}\big[\log \hat{P}_j\big]} \qquad \forall\ 1 \le i \le C.$$

(Hint: To implement the positivity constraint on R, you can reparameterize its components as $R_i = \exp(Z_i)$, and minimize the objective w.r.t. Z.)

Solution:

$$\min_R \mathrm{E}[D_{KL}(R\|\hat{P})] = \min_R \mathrm{E}\left[\sum_{i=1}^{C} R_i \log\left(\frac{R_i}{\hat{P}_i}\right)\right]$$

If we replace the R_i with $R_i = \exp(Z_i)$ and consider the fact that R_i is a probability, then we can reformulate the optimization

$$\min_Z \mathrm{E}\left[\sum_{i=1}^{C} \exp(Z_i)\left(\log(\exp(Z_i)) - \log \hat{P}_i\right)\right]$$

$$\text{s.t.} \quad \sum_{i=1}^{C} \exp(Z_i) = 1$$

$$\Downarrow$$

$$\min_Z \mathrm{E}\left[\sum_{i=1}^{C} \exp(Z_i) Z_i - \exp(Z_i) \log \hat{P}_i\right]$$

$$\text{s.t.} \quad \sum_{i=1}^{C} \exp(Z_i) = 1$$

爽爺出

$$\min_{Z} \sum_{i=1}^{C} \exp(Z_i) Z_i - \exp(Z_i) E[\log \hat{P}_i]$$

$$\text{s.t.} \quad \sum_{i=1}^{C} \exp(Z_i) = 1$$

Then we can use the lagrange multiplier to solve this constrainted optimization problem. ✓

$$\mathcal{L}(Z, \lambda) = \sum_{i=1}^{C} \exp(Z_i) Z_i - \exp(Z_i) E[\log \hat{P}_i] + \lambda(\sum_i \exp(Z_i) - 1)$$

compute the derivatives w.r.t Z_i and λ

$$\frac{\partial \mathcal{L}}{\partial Z_i} = \exp(Z_i)(Z_i + 1) - \exp(Z_i) E[\log \hat{P}_i] + \lambda \exp(Z_i)$$

$$= \exp(Z_i)(Z_i + 1 + \lambda) - \exp(Z_i) E[\log \hat{P}_i] = 0$$

Since $\exp(Z_i) > 0$, then: ✓

$$Z_i + 1 + \lambda = E[\log \hat{P}_i] \longrightarrow Z_i = E[\log \hat{P}_i] - 1 - \lambda$$

$$\therefore R_i = \exp(Z_i) = \exp(E[\log \hat{P}_i]) / \exp(1 + \lambda)$$

Since R_i is a probability density, then we have ✓

$$\sum_{i=1}^{C} R_i = \frac{\sum_{i=1}^{C} \exp(E[\log \hat{P}_i])}{\exp(1 + \lambda)} = 1$$

which means that :

$$\exp(1+\lambda) = \sum_{i=1}^{C} \exp(E[\log \hat{P}_i])$$

$$\therefore \quad R_i = \frac{\exp(E[\log \hat{P}_i])}{\sum_{i=1}^{C} \exp(E[\log \hat{P}_i])} \qquad \forall i = 1 \cdots C$$

proofed.

(b) *Prove* the bias-variance decomposition

$$\text{Error}(\hat{P}) = \text{Bias}(\hat{P}) + \text{Var}(\hat{P})$$

where the error, bias and variance are given by

$$\text{Error}(\hat{P}) = \mathrm{E}\big[D_{\mathrm{KL}}(P||\hat{P})\big], \qquad \text{Bias}(\hat{P}) = D_{\mathrm{KL}}(P||R), \qquad \text{Var}(\hat{P}) = \mathrm{E}\big[D_{\mathrm{KL}}(R||\hat{P})\big].$$

(Hint: as a first step, it can be useful to show that $\mathrm{E}[\log R_i - \log \hat{P}_i]$ *does not depend on the index* i*.)*

Solution:

From (a) we have :

$$R_i = \frac{\exp(E[\log \hat{P}_i])}{\sum_{i=1}^{C} \exp(E[\log \hat{P}_i])}$$

① First proof $E[\log R_i - \log \hat{P}_i]$ doesn't depend on index i.

$$E[\log R_i - \log \hat{P}_i] = E[E[\log \hat{P}_i] - \log(\sum_{i=1}^{C} \exp(E[\log \hat{P}_i]) - \log \hat{P}_i] \quad (M)$$

$$= E[\log \hat{P}_i] - E[\log \hat{P}_i] - E[\log M]$$

$$= -E_{(R_i)}[\log M]$$

$$= -\log M$$

Since M is the sum of $i \in \{1, \cdots, C\}$

$\therefore E[\log R_i - \log \hat{P}_i]$ is independent of the index.

② Then prove the bias-variance decomposition.

$$\text{Error}(\hat{P}) = E[D_{KL}(P \| \hat{P})]$$

$$= E\left[\sum_i P_i \log P_i - P_i \log \hat{P}_i + P_i \log R_i - P_i \log R_i\right]$$

$$= E\left[\sum_i P_i \log P_i - P_i \log R_i\right] + E\left[\sum_i -P_i \log \hat{P}_i + P_i \log R_i\right]$$

$$= D_{KL}(P \| R) + \sum_i P_i \underbrace{E[\log R_i - \log \hat{P}_i]}_{\text{independant of index } i}$$

$$= D_{KL}(P \| R) + E[\log R_i - \log \hat{P}_i] \cdot \left(\sum_i P_i\right)$$

since P_i is a density function as the same as R_i

$\therefore \quad \Sigma_i P_i = \Sigma_i R_i = 1$

$$= D_{KL}(P \| R) + E[\log R_i - \log \hat{P}_i] \cdot (\Sigma_i R_i)$$

$$= D_{KL}(P \| R) + E[\Sigma_i R_i \log R_i - R_i \log \hat{P}_i]$$

$$= D_{KL}(P \| R) + E[D_{KL}(R \| \hat{P})]$$

$$= \text{Bias}(\hat{P}) + \text{Var}(\hat{P})$$

proofed

Index der Kommentare