Machine Learning I at TU Berlin

Assignment 7 - Group PTHGL

December 3, 2018

Exercise 1

Bias and Variance of Mean Estimators
The following formulas give the bias, variance and mean squared error for an estimator $\hat{\theta}$ for a parameter θ.

$$\begin{aligned}
\text{Bias}(\hat{\theta}) &= \text{E}[\hat{\theta} - \theta] \\
\text{Var}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \text{E}(\hat{\theta}))^2] \\
\text{Error}(\hat{\theta}) &= (\text{Bias}(\hat{\theta}))^2 + \text{Var}(\hat{\theta}) = (\text{E}[\hat{\theta} - \theta])^2
\end{aligned}$$

(a)

For the estimator $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ we calculate the bias, variance and error:

$$\begin{aligned}
\text{Bias}(\hat{\theta}) = \text{E}[\hat{\mu} - \mu] &= \text{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right] \\
&= \text{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] - \mu \quad | \quad \text{linearity of E} \\
&= \frac{1}{N}\sum_{i=1}^{N} \underbrace{\text{E}[X_i]}_{=\mu} - \mu \\
&= \frac{1}{N} N\mu - \mu \\
&= 0
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \quad | \quad \mathrm{Var}(cX) = c^2\mathrm{Var}(X) \\
&= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) \quad | \quad \text{samples are i.i.d. } (\textit{Bienaym formula}) \\
&= \frac{1}{N^2}\sum_{i=1}^{N} \underbrace{\mathrm{Var}(X_i)}_{=\sigma^2} \\
&= \frac{1}{N^2} N\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned}
$$

With the bias and the variance we can calculate the mean squared error:

$$
\begin{aligned}
\mathrm{Error}(\hat{\theta}) &= (\mathrm{Bias}(\hat{\theta}))^2 + \mathrm{Var}(\hat{\theta}) = 0^2 + \frac{\sigma^2}{N} \\
&= \frac{\sigma^2}{N}
\end{aligned}
$$

(b)

For the estimator $\hat{\mu} = 0$ we calculate the bias, variance and error:

$$
\begin{aligned}
\mathrm{Bias}(\hat{\theta}) &= \mathrm{E}[\hat{\mu} - \mu] = \mathrm{E}[0 - \mu] \\
&= \mathrm{E}[0] - \mathrm{E}[\mu] = 0 - \mu \\
&= -\mu \\
\mathrm{Var}(\hat{\theta}) &= \mathrm{E}[(\hat{\mu} - \mathrm{E}[\hat{\mu}])^2] \\
&= \mathrm{E}[(0 - \mathrm{E}[0])^2] \\
&= 0
\end{aligned}
$$

With the bias and the variance we can calculate the mean squared error:

$$
\begin{aligned}
\mathrm{Error}(\hat{\theta}) &= (\mathrm{Bias}(\hat{\theta}))^2 + \mathrm{Var}(\hat{\theta}) = (-\mu)^2 + 0 \\
&= \mu^2
\end{aligned}
$$

Exercise 2

Bias-Variance Decomposition for Regression
The function $y = f(x)$ is mapping from input to output. The estimator $\hat{f}(x)$ is obtained be training a regression model on some random sample $D = (x_1, y_1), ..., (x_N, y_N)$ of the mapping

$y = f(x)$. The bias, variance and error is given by

$$\text{Error}(\hat{f}(x)) = \text{E}\left[(\hat{f}(x) - f(x))^2\right], \quad \text{Var}(\hat{f}(x)\text{E}\left[(\hat{f}(x) - \text{E}[\hat{f}(x)])^2\right], \quad \text{Bias}(\hat{f}(x) = \text{E}\left[(\hat{f}(x) - f(x))\right].$$

In the following, we prove, that $\text{Error}(\hat{f}(x)) = \text{Var}(\hat{f}(x) + \text{Bias}(\hat{f}(x))^2$:

$$\begin{aligned}
\text{Error}(f(x) &= \text{E}[\hat{f}(x)^2 - 2f(x)\hat{f}(x) + f(x)^2] \\
&= \text{E}[\hat{f}(x)^2] - 2\text{E}[f(x)\hat{f}(x))] + f(x))^2 \\
&= \text{E}[\hat{f}(x))^2] - 2\text{E}[\hat{f}(x)]^2 + \text{E}[\hat{f}(x)]^2 + \text{E}[\hat{f}(x)]^2 - 2f(x)\text{E}[\hat{f}(x)] + f(x)^2 \\
&= \text{E}[\hat{f}(x))^2] - 2\text{E}[\hat{f}(x)\text{E}[\hat{f}(x)]] + \text{E}[\text{E}[\hat{f}(x)]^2] + \text{E}[\hat{f}(x)]^2 - 2f(x)\text{E}\hat{f}(x)] + f(x)^2 \\
&= \text{E}[(\hat{f}(x)^2 - \text{E}[\hat{f}(x)])^2] + (E[\hat{f}(x) - f(x))^2 \\
&= \text{Var}(\hat{f}(x) + \text{Bias}(\hat{f}(x))^2
\end{aligned}$$

Exercise 3

Bias-Variance Decomposition for Classication

(a)

Given the optimization problem

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

with the class distribution estimator $\hat{P} = \{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_C\}$ and the distribution $R = \{R_1, R_2, \ldots, R_C\}$, we derive the discrete KL divergence

$$D_{KL} = \sum_{i=1}^{C} R_i \log\left(\frac{R_i}{\hat{P}_i}\right) = \sum_{i=1}^{C} R_i \log(R_i) - \sum_{i=1}^{C} R_i \log\left(\hat{P}_i\right)$$

for one distribution R_i. Afterwards we normalize the solution, so that

$$\sum_i R_i = 1$$

.

$$\min_{R_i} \text{E}[D_{KL}(R||\hat{P})] = \text{E}\left[\min_{R_i} D_{KL}(R||\hat{P})\right]$$

Minimize D_{KL}:

$$\begin{aligned}
\frac{\partial D_{KL}}{\partial R_i} &= (\log R_i + 1) - \log \hat{P}_i = 0 \\
&\rightarrow \log R_i = \log \hat{P}_i - 1
\end{aligned}$$

Expected value $\mathrm{E}\left[\min_{R_i} D_{KL}(R||\hat{P})\right]$:

$$\mathrm{E}[\log R_i] = \mathrm{E}[\log \hat{P}_i - 1] = \mathrm{E}[\log \hat{P}_i] - \underbrace{\mathrm{E}[1]}_{=1}$$

$$\mathrm{E}[\log R_i] = \sum_{j=1}^{C} \hat{P}_j \log R_i = \mathrm{E}[\log \hat{P}_i] - 1 \qquad | \sum_{j=1}^{C} \hat{P}_j = 1$$

$$\log R_i = \mathrm{E}[\log \hat{P}_i] - 1 \qquad | \cdot \exp\{\}$$

$$R_i = \exp\Big\{\mathrm{E}[\log \hat{P}_i] - 1\Big\} = \exp\Big\{\mathrm{E}[\log \hat{P}_i]\Big\} \cdot \exp\{-1\}$$

$$R_i = \frac{\exp\Big\{\mathrm{E}[\log \hat{P}_i]\Big\}}{\exp\{1\}}$$

Normalization with

$$\frac{R_i}{\sum_{j=1}^{C} R_j}$$

$$\sum_{j=1}^{C} R_j = \sum_{j=1}^{C} \frac{\exp\Big\{\mathrm{E}[\log \hat{P}_j]\Big\}}{\exp\{1\}}$$

$$\begin{aligned} R_{i(norm)} &= \frac{R_i}{\sum_{j=1}^{C} R_j} \\ &= \frac{\exp\Big\{\mathrm{E}[\log \hat{P}_i]\Big\}}{\cancel{\exp\{1\}}} \cdot \sum_{j=1}^{C} \frac{\cancel{\exp\{1\}}}{\exp\Big\{\mathrm{E}[\log \hat{P}_j]\Big\}} \\ &= \frac{\exp\Big\{\mathrm{E}[\log \hat{P}_i]\Big\}}{\sum_{j=1}^{C} \exp\Big\{\mathrm{E}[\log \hat{P}_j]\Big\}} \end{aligned}$$

(b)

The bias, variance and error is given by

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

In the following, we prove $\text{Error}(\hat{P}) = \text{Bias}(\hat{P}) + \text{Var}(\hat{P})$.

$$
\begin{aligned}
\text{Error}(\hat{P}) &= \text{E}[D_{KL}(P||\hat{P})] = \sum_{i=1}^{C} P_i \log\left(\frac{P_i}{\hat{P}_i}\right) \\
&= \text{E}[\text{E}[\log P - \log \hat{P}]] \\
&= \text{E}[\text{E}[\log P - \log R + \log R - \log \hat{P}]] \\
&= \text{E}[\log P - \log R] + \text{E}[\text{E}[\log R - \log \hat{P}]] \\
&= \sum_{i=1}^{C} P_i \log\left(\frac{P_i}{R_i}\right) + \text{E}\left[\sum_{i=1}^{C} P_i \log\left(\frac{R_i}{\hat{P}_i}\right)\right] \\
&= D_{KL}(P||R) + \text{E}[D_{KL}(R||\hat{P})] \\
&= \text{Bias}(\hat{P}) + \text{Var}(\hat{P})
\end{aligned}
$$

Final Solutions

December 3, 2018

1 Model Selection

In this programming assignment we examine techniques for model selection on classification and regression tasks. In particular, we first explore the effect of model hyperparameters on the bias and variance of the prediction. In the second part of the assignment we utilize the bias-variance decomposition to perform automatic hyperparameter selection. Several classes and methods are provided in the `utils.py` file:

1.0.1 Datasets

- `utils.Housing()`: This regression dataset is available at http://archive.ics.uci.edu/ml/datasets/Housing and loaded from scikit-learn's inbuilt representation. This data is used for regression. A description of the dataset can be found here http://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names. This data is in a `506x13` matrix and the labels in a array of length `506`.

- `utils.Yeast()`: This classification dataset is available at https://archive.ics.uci.edu/ml/datasets/Yeast . This data is used for classification. A description of the dataset can be found here https://archive.ics.uci.edu/ml/machine-learning-databases/yeast/yeast.names. This data is in a `1484x8` matrix and the labels (class probabilities) are in a `1484x7` matrix where `targets[i,j] = 1` if example `i` is of class `j` and `0` otherwise. For example, if we have a dataset of 4 examples which belong to following classes : `[1, 0, 0, 2]` the label matrix would look like this: `T = [[0, 1, 0], [1, 0, 0], [1, 0, 0], [0, 0, 1]]`.

1.0.2 Predictors

We provide two simple classes of predictors, one for regression and one for classification:

- `utils.ParzenRegression`: 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.ParzenClassification`: 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`.

1.0.3 Bias Variance Decomposition

As we have seen in the theoretical exercise, there are several possible bias-variance decomposition for different tasks (e.g. classification, or regression).

- `utils.biasVarianceRegression()`: Perform the usual bias-variance decomposition of the mean square error. Reminder: given Y the (random) estimator and T the target, the decomposition is computed as follows:
- $\text{Bias}(Y)^2 = (\mathbb{E}_Y[Y - T])^2$
- $\text{Var}(Y) = \mathbb{E}_Y[(Y - \mathbb{E}_Y[Y])^2]$
- $\text{Error}(Y) = \mathbb{E}_Y[(Y - T)^2]$

1.0.4 Sampler

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.

1.1 Part 1: Implementing Bias-Variance Decomposition for Classification (20 P)

Implement a function which computes the bias, variance and error given the true labels of the training data and the predicted values. Bias, Variance and Error for classification are defined as:

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

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. Note that we consider here the Kullback-Leibler divergence as a measure of classification error, which is commonly done in practice in order to have a smooth objective function.

Tasks:

- **Implement the KL-based Bias-Variance Decomposition defined above (10 P)**

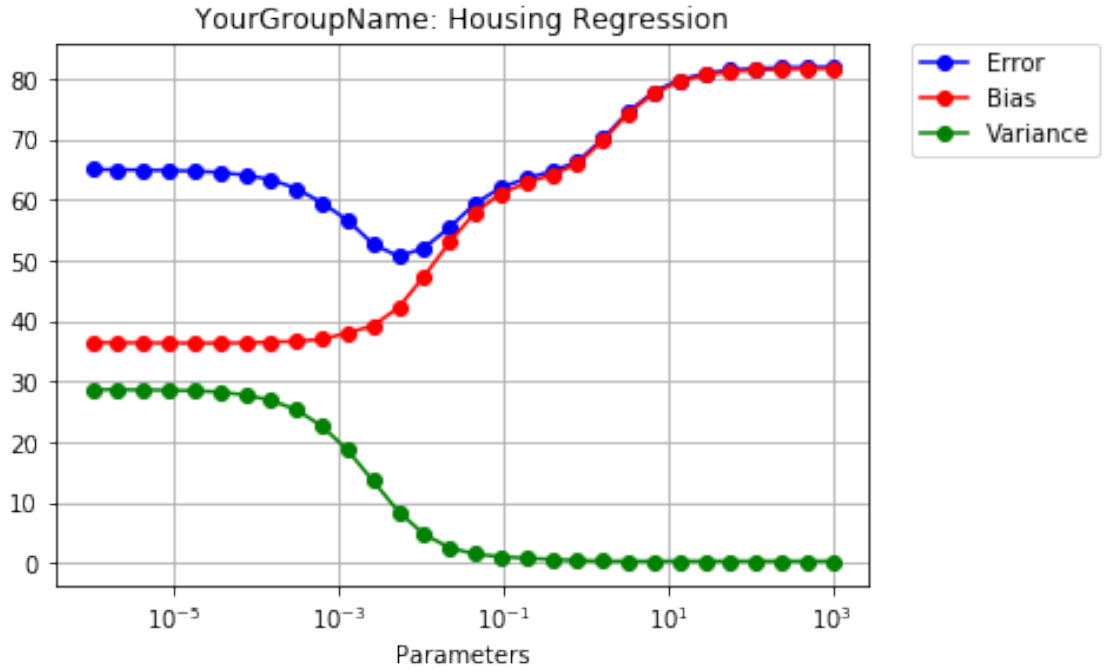
To get started, you can take inspiration from the readily implemented function `utils.biasVarianceRegression()`, which does the following:

- Iterate for a certain number of times the following:
  - Acquire a subsample of the training data by invoking `sampler.sample()`
  - Using the predictor (which will either be a Parzen Regressor or Parzen Classifier depending on the task), fit the model on the sample and determine the prediction for the observation data (N examples disjoint from the training data). Note that the dimension of the outputs matches the dimension of the targets, so for regression you will get an array of length N and for classification a matrix of shape $N \times$ #classes containing the class distributions.
- Having computed a number of different predictions, determine the bias, variance and error comparing the predictions to the true labels. Check that the decomposition is correct (i.e. bias + variance = error) using an `assert` statement, and return the bias and variance.
- **Once the method is implemented, run Test 1 and Test 2 provided below (10 P)**

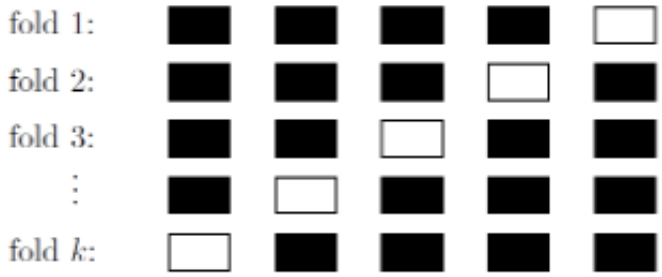
```
In [1]: import numpy as np

        def biasVarianceClassification(sampler, predictor, X, T, nbsamples=25):
            Y = np.array([predictor.fit(*sampler.sample()).predict(X) for _ in range(nbsamples)]

            Ri = np.exp(np.mean(np.log(Y), axis=0))
            Risum = np.sum(Ri, axis=1)
            R = np.zeros([Ri.shape[0],Ri.shape[1]])
            #R[:,Ri.shape[1]] = Ri[:,Ri.shape[1]]/ Risum
            for i in range(Ri.shape[1]):
                R[:,i] = Ri[:,i] / Risum

            #bias
            DTR = np.sum(T * np.log(T / R),axis=1)
            bias = np.mean(DTR, axis=0)

            #Variance
            DRY = np.zeros([Y.shape[0],Y.shape[1],Y.shape[2]])
            for i in range(Y.shape[2]):
                DRY[i,:,:] = R * np.log(R / Y[i,::])
            #DRY[Y.shape[2],:,:] = R * np.log(R / Y[Y.shape[2],::])
            variance = np.mean(np.mean(np.sum(DRY, axis=2),axis=0))

            #Error
            DTY = np.zeros([Y.shape[0],Y.shape[1],Y.shape[2]])
            for i in range(Y.shape[2]):
                DTY[i,:,:] = T * np.log(T / Y[i,::])
            #DTY[Y.shape[2],:,:] = T * np.log(T / Y[Y.shape[2],::])
            error = np.mean(np.mean(np.sum(DTY, axis=2),axis=0))
```

```
        #assert(numpy.abs((bias + variance) / error - 1) < 1e-4)

        return bias, variance
```

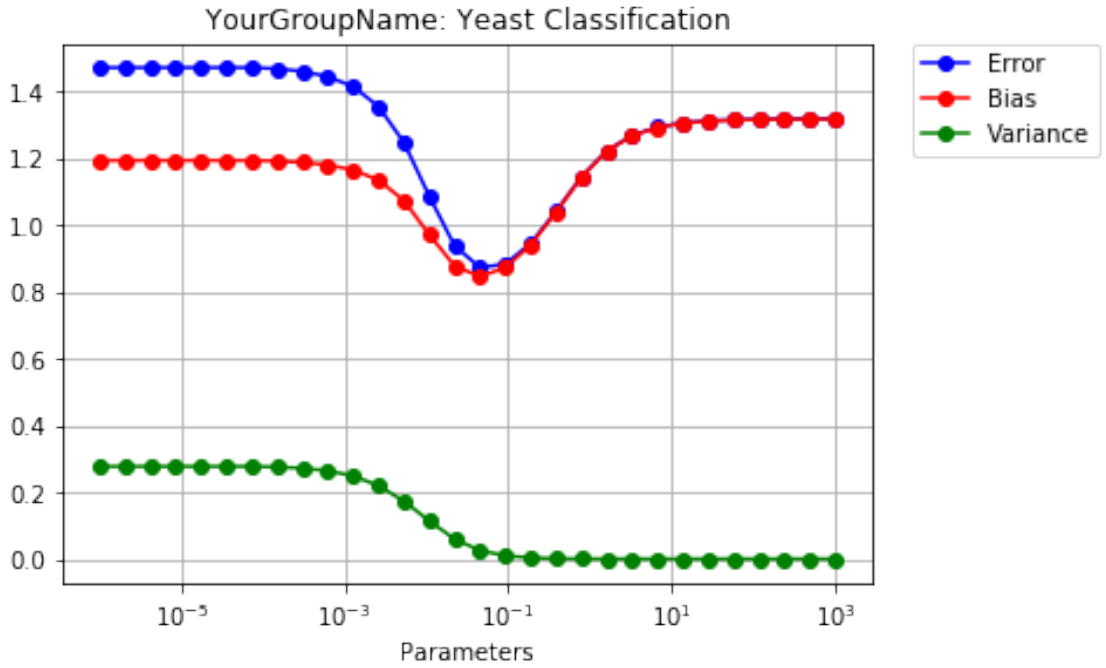
In [2]:
```
### TEST 1
import utils,numpy
%matplotlib inline
utils.plotBVE(utils.Housing,numpy.logspace(-6,3,num=30),utils.ParzenRegressor,utils.bias
```



In [3]:
```
### TEST 2
import utils,numpy
%matplotlib inline
utils.plotBVE(utils.Yeast,numpy.logspace(-6,3,num=30),utils.ParzenClassifier,biasVarianc
```



Loop



1.2 Part 2: Implementing a Parameter Selection Procedure (30 P)

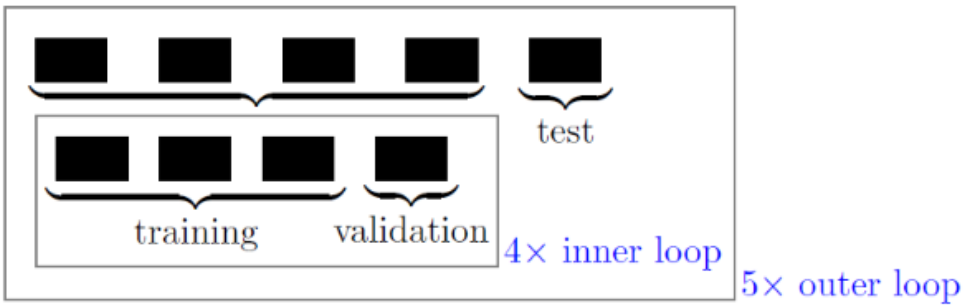
In this part of the exercise, we would like to find what is the best hyperparameter of the model for predicting the Housing regression data. A 5-fold cross-validation procedure is already implemented and that allows to compute error bars.

You need to extend this basic cross-validation procedure by a nested loop of 4-fold cross-validation that selects the best hyperparameters based on some criterion (cost function) to be determined. The nested loop is depicted below:

The full procedure for evaluation and hyperparameter selection procedure is shown in the diagram below with the part that you need to implement highlighted in red.

Tasks:

- **Implement the inner loop of 4-fold cross-validation, helping you from the diagram above**



Nested

Build several **selection** / **test** splits from the data

Loop over **selection** / **test** splits

- Build several **training** / **validation** splits from the **selection** set
- Loop over **training** / **validation** splits
  - Loop over model parameters
    - Train several models with various random samples of the **training** set
    - Compute the bias and variance of these models on the **validation** set
    - Compute *cost*(bias,variance)
  - Return the best parameter for this split (i.e. with the smallest associated cost)
- Use as parameter the *geometric* mean of best parameters for each split
- Train several models with various random samples of the **selection** set
- Compute the bias and variance of these models on the **test** set

Compute error bars (average and standard deviation) for bias, variance and error.

Procedure

(20 P)

For this part, use the following settings:

- Range of parameters to test: 15 parameters logarithmically spaced between 1e-5 and 1e5.
- The returned parameter is the geometric mean of the best parameter found for each split.
- The best parameter for each split is the one that minimizes the `costfunction` specified as argument.
- The bias and variance estimates are obtained by sampling 10 times from the training distribution.
- **Verify your implementation by running Test 3 (10 P)**

```
In [4]: def getbestparameter(Xselect,Tselect,costfunction):
            splits = [([1,2,3],0) , ([0,2,3],1) , ([0,1,3],2) , ([0,1,2],3)]
            nbsamples = 10
            finalbestparam = 1

            #Loop over selection/test splits
            for inds_train,ind_valid in splits:

                Xtrain = [Xselect[ind] for ind in inds_train]
                Ttrain = [Tselect[ind] for ind in inds_train]

                Xvalid = Xselect[ind_valid]
                Tvalid = Tselect[ind_valid]

                bestcost = 1000000
                bestparam = 0;
                for param in np.logspace(-5,5,num=15):
                    predictor = utils.ParzenRegressor(param)
                    sampler = utils.Sampler(np.concatenate(Xtrain,axis=0),np.concatenate(Ttrain,
                    bias,variance = utils.biasVarianceRegression(sampler,predictor, Xvalid,Tvali
                    cost = costfunction(bias, variance)

                    if(cost < bestcost):
                        bestcost = cost
                        bestparam = param
            return bestparam

In [5]: import numpy,utils

        def evaluateModel(X,T,costfunction):
            # X: partitioned input
            # T: partitioned targets
            # costfunction: the function for evaluate how good/bad a hyperparameter is
```

```
    # Create splits
    splits = [ ([1,2,3,4],0) , ([0,2,3,4],1) , ([0,1,3,4],2) , ([0,1,2,4],3) , ([0,1,2,3

    testbiases,testvariances,testerrors,bestparameters = [],[],[],[]

    #Loop over selection/test splits
    for inds_select,ind_test in splits:

        Xselect = [X[ind] for ind in inds_select]
        Tselect = [T[ind] for ind in inds_select]

        Xtest = X[ind_test]
        Ttest = T[ind_test]

        bestparam = getbestparameter(Xselect,Tselect,costfunction)

        # Evaluate bias and variance with this best parameter
        predictor = utils.ParzenRegressor(bestparam)
        sampler   = utils.Sampler(numpy.concatenate(Xselect,axis=0),numpy.concatenate(Ts
        bias,variance = utils.biasVarianceRegression(sampler,predictor,Xtest,Ttest, nbsa

        testbiases     += [bias]
        testvariances  += [variance]
        testerrors     += [bias+variance]
        bestparameters += [bestparam]

    # Output results of model evaluation
    print('bias:      %8.5f +/- %8.5f'%(numpy.mean(testbiases),numpy.std(testbiases)))
    print('variance:  %8.5f +/- %8.5f'%(numpy.mean(testvariances),numpy.std(testvariance
    print('error:     %8.5f +/- %8.5f'%(numpy.mean(testerrors),numpy.std(testerrors)))
    print('parameter: %8.5f +/- %8.5f'%(numpy.mean(bestparameters),numpy.std(bestparamet
```

In [6]:
```
### TEST 3

import numpy,utils

costfunctions = [
         ('Parameter Selection Criterion: favor low bias', lambda b,v: 9*b+v),
         ('Parameter Selection Criterion: favor low error',lambda b,v: b+v),
         ('Parameter Selection Criterion: favor low variance',lambda b,v: b+9*v),
]

# Load and partition the data
X,T = utils.Housing()
n = len(X)
X = [X[n*i//5:n*(i+1)//5] for i in range(5)]
T = [T[n*i//5:n*(i+1)//5] for i in range(5)]
```

```
        print ("YourGroupName")
        for name,costfunction in costfunctions:
            print('\n\n%s\n'%name)
            evaluateModel(X,T,costfunction)
```

```
YourGroupName


Parameter Selection Criterion: favor low bias

bias:      38.72012 +/-  6.51560
variance:  18.66858 +/-  4.96238
error:     57.38870 +/-  6.35303
parameter:  0.00034 +/-  0.00054


Parameter Selection Criterion: favor low error

bias:      42.62598 +/-  7.68426
variance:   6.50808 +/-  4.27666
error:     49.13406 +/-  7.23962
parameter:  0.00487 +/-  0.00285


Parameter Selection Criterion: favor low variance

bias:      60.14089 +/- 10.53673
variance:   0.92393 +/-  0.23890
error:     61.06482 +/- 10.51482
parameter:  0.06843 +/-  0.06232


In [ ]:
```