Danmarks
Tekniske
Universitet

DTU

Project 2

02450 Introduction to Machine Learning and Data Mining

Group 7

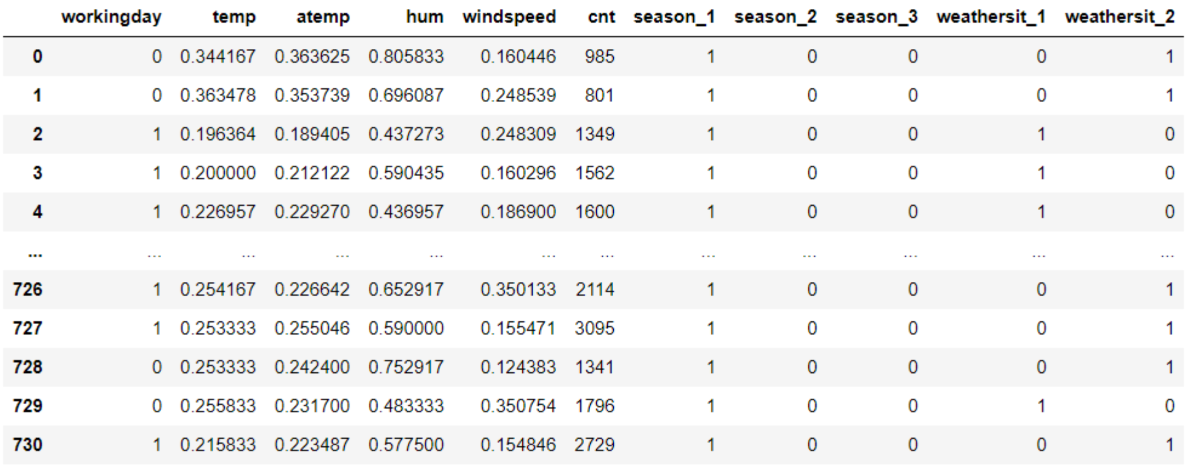
Student Contribution

Student Name - ID	Regression part a	Regression part b	Classification	Exam Questions
Yu Fan Fong - s230003	40%	100%	65%	0%
Ashish Rakesh Chandra Kukreti - s230134	0%	0%	35%	0%
Karrar Adam Mahdi - s230432	60%	0%	0%	100%

Regression, part a:

The daily total number of bike rental users (cnt) is the target variable based on other variables like working day, season, weather situations, temperature, humidity, wind speed, and apparent temperature. The aim of the regression model is to assist bike rental operators in anticipating bike rental demand and prepare their fleet accordingly. In the dataset, each row corresponds to a day over a 2-year period.

One-hot encoding was applied to the categorical variables (working day, season and weather conditions). The 4 seasons were represented using 3 binary variables (season_1, season_2, season_3), such that each binary variable represents the difference between season_1 and season_4. Also, if season_1=season_2=season_3=0, it can be interpreted as season_4=1. Similarly, the 3 weather situations were represented using 2 binary variables. To remove the effects of the different scales of the variables, the continuous variables were standardised so that the relative importance of the factors become more apparent. After the initial test runs, the data for the number of casual and registered bike users were dropped from the dataset as the total count variable is directly related to the sum of casual and registered users. Thus, including casual and registered users resulted in high weights for these two variables only. Hence, they were removed to produce more meaningful results.



	workingday	temp	atemp	hum	windspeed	cnt	season_1	season_2	season_3	weathersit_1	weathersit_2
0	0	0.344167	0.363625	0.805833	0.160446	985	1	0	0	0	1
1	0	0.363478	0.353739	0.696087	0.248539	801	1	0	0	0	1
2	1	0.196364	0.189405	0.437273	0.248309	1349	1	0	0	1	0
3	1	0.200000	0.212122	0.590435	0.160296	1562	1	0	0	1	0
4	1	0.226957	0.229270	0.436957	0.186900	1600	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
726	1	0.254167	0.226642	0.652917	0.350133	2114	1	0	0	0	1
727	1	0.253333	0.255046	0.590000	0.155471	3095	1	0	0	0	1
728	0	0.253333	0.242400	0.752917	0.124383	1341	1	0	0	0	1
729	0	0.255833	0.231700	0.483333	0.350754	1796	1	0	0	1	0
730	1	0.215833	0.223487	0.577500	0.154846	2729	1	0	0	0	1

Figure 1. Preview of the transformed dataset used for training.

In an attempt to improve the generalization performance of the regression model, regularisation was applied to the transformed dataset. Better predictions on the demand of bike rentals could be made as the trained model might have better generalisation with less overfitting to the training data. The regularised linear regression model used λ values in the range (10^{-5}, 10^{9}) with K=10 fold cross-validation as in the figure 2 below.

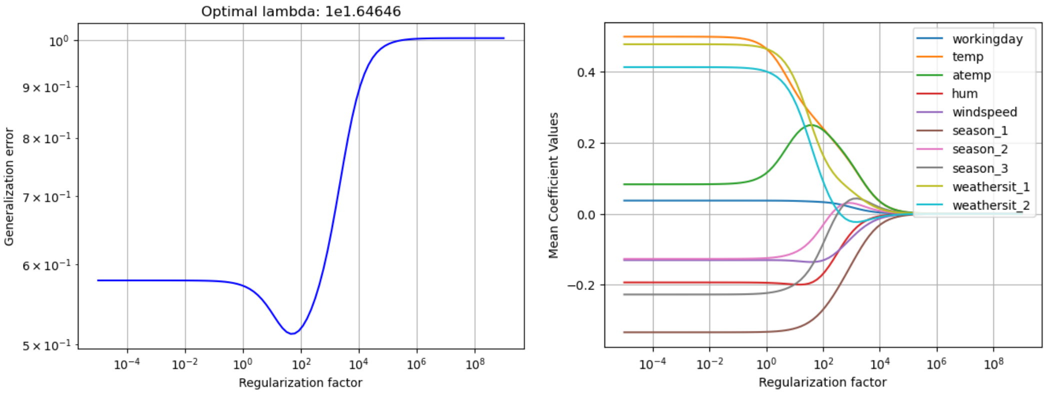


Figure 2. (left) Graph of estimated generalisation error against regularisation factor. (right) Graph of mean coefficient values against regularisation factor.

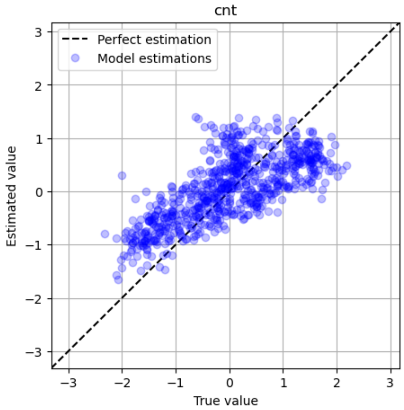


Figure 3. Graph of estimated values against true value of linear regression. The regularised linear regression model was trained on the entire dataset with the optimal lambda found to produce this visualisation.

When the regularization λ is small, the models have high variance and low bias. As λ increases, the model variance decreases, while the bias increases, and the weights all dragged towards the x-axis (figure 2). Thus, it is expected that the high bias at large λ values result in greater generalisation errors. The generalisation error dropped from $\lambda=10^0$, reaching a minimum at $\lambda = 10^{1.65}$, and then increased afterwards. The optimal lambda found is $10^{1.65}$ (figure 2). The decrease in generalisation error is not significant, with a decrease about less than 0.1. This is likely due to the weights being relatively small to begin with ($|w|<0.6$), thus regularisation made smaller modifications to weights and had an overall smaller effect on the generalisation error. Hence, regularisation may not be needed for this regression problem. The overall performance of the regularised linear regression model with optimal λ used can be seen in figure 3.

From figure 2, the top two most significant factors are the temperature and weathersit_1 (clear/partly cloudy). Low values of season_1 (winter), season_3 (summer), humidity and high values of temperature, weathersit_1, and weathersit_2 (mist) lead to high predictions of total count of bike users. Factors like whether it is a working day, windspeed, season_2 (spring) and a_temp (apparent temperature) are less significant. This is surprising because in consideration of the outdoor weather experienced by bike users, they are physically experiencing the apparent temperature and wind speed, but temp is a stronger indicator than a_temp and wind speed is not a strong indicator of the total count. Again, it is contradicting that both the winter and summer seasons have strong negative weights. One would expect summer to have more bike users due to the warmer outdoors. Hence, this suggests that the total count of bike users goes beyond looking at the outdoor environment and it may be worth it to consider more factors that are not weather-related.

Regression, part b:

The regularised linear regression model used λ values in the range (10^{-5}, 10^{9}). At the start, the ANN model was trained with the number of hidden units (***h*** values) between 1 and 10, but it generally favoured the smaller number of hidden units. Hence, the final set of ***h*** values used was reduced to [2,4,6]. The maximum iterations was increased to 20000, while the number of replicates trained was 1 due to time constraint. The baseline model computes the mean of the training data and uses it to predict y on the test data. For the 2-level cross validation, $K_1 = K_2 =$ 10 was used. The same D_{par}^{i} ,D_{test}^{i} splits were used to retrain the 3 methods. Mean squared error was used as the cost function.

	ANN		**Linear Regression**		**Baseline**
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	2	0.472	1e0.939	0.542	0.953
2	2	0.447	1e0.939	0.612	0.979
3	2	0.419	1e0.939	0.475	1.098
4	4	0.292	1e0.939	0.350	1.137
5	4	0.391	1e0.939	0.459	0.950
6	4	0.307	1e0.939	0.344	0.705
7	4	0.469	1e0.939	0.469	0.985
8	4	0.386	1e0.939	0.459	1.141
9	4	0.463	1e0.939	0.520	0.988
10	2	0.382	1e0.939	0.450	0.895

Table 1. Two-level cross-validation table used to compare the three models for regression, rounded to 3dp.

The regularised linear regression and ANN models consistently had lower test errors than the baseline model. The optimal λ values found are lower than that obtained in part a. Again, the decrease in test error is also minimal due to the relatively small weights to begin with (figure 4). From figure 4, the top two most significant factors are still the temperature and weathersit_1 (clear/partly cloudy).

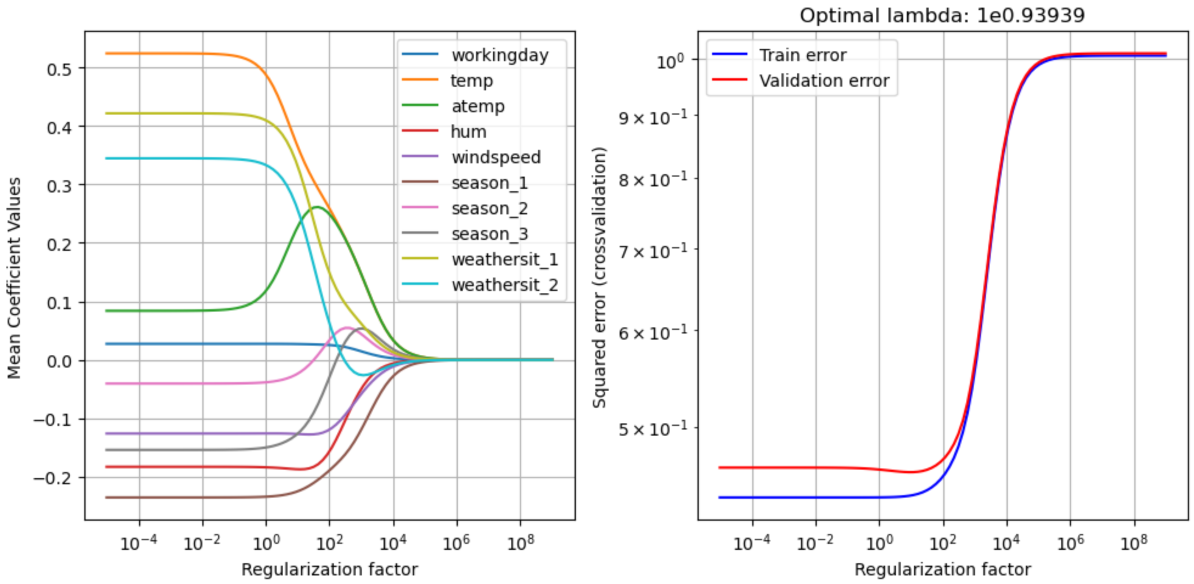


Figure 4. (left) Graph of mean weight values against regularisation factor. (right) Graph of mean squared error of linear regression against regularisation factor.

Statistical Evaluation

Statistical evaluation was carried out following setup I using paired t-tests at a α=0.05 significance level. Thus, conclusions established in the following analysis are dependent on this specific bike sharing dataset. The confidence intervals (CI) are indications of better performance of model A against model B. If the CI is below zero, it means that there is 1-α probability that model A is better.

Model A	ANN	ANN	Regularised Linear Regression
Model B	Regularised Linear Regression	Baseline	Baseline
CI	(-0.0976, -0.0347)	(-0.659, -0.499)	(-0.590, -0.436)
p-value	4.14e-05	2.63e-40	1.92e-35
Performance Conclusion	$M_{\text{ANN}} > M_{\text{Reg Linear Regression}}$	$M_{\text{ANN}} > M_{\text{Baseline}}$	$M_{\text{Reg Linear Regression}} > M_{\text{Baseline}}$

Table 2. Confidence intervals and p-values obtained from comparing the 3 methods pairwise, rounded to 3sf.

Since all 3 p-values are well below the 0.05 significance level, the results observed are unlikely due to chance, so it is unlikely that the three models have the same performance. From the confidence intervals, it can be inferred that the performance of $M_{\text{ANN}} > M_{\text{Reg Linear Regression}} > M_{\text{Baseline}}$ when trained on this specific bike sharing dataset at the significance level of 0.05. Hence, it can be concluded that the ANN and regularised linear regression models have been able to produce meaningful learning outcomes and are able to provide better predictions than a trivial baseline model.

According to these findings, the ANN model is the best model for the regression problem of this dataset. However, the regularised logistic regression model still provides decent predictions and requires much less computational resources. Thus, the choice of model may be influenced by other factors like processing resources and accuracy requirements of the bike rental operator.

Classification:

The classification problem chosen for this project is to predict whether a day is a working day or not. This is a binary classification problem whereby the target variable has two classes: ***'yes'*** or ***'no'***. The goal is to create a model that can reliably predict whether a day is a working day or not.

Casual and registered users were included back into the dataset as these numbers may vary depending on the work day. For instance, registered users may rely on the bikes as their mode of transport to work, while casual users may use the bikes while they are taking a break on a non-working day.

In this study, three machine learning methods were used: regularised logistic regression, K-nearest neighbours (KNN), and a baseline model. To control the model complexity, **λ** and ***k*** were used as complexity-controlling parameters for the logistic regression and KNN models respectively. The KNN model used ***k*** values from 1 to 100, while the regularised logistic regression model used **λ** values in the range (10^{-5}, 10^{9}). Logistic regression seeks to compute the feature coefficients that best splits the data into two classes (working day or not). It was implemented using Scikit-learn's linear model LogisticRegression. KNN is an algorithm for classification that predicts the class of a given input data point based on its feature space nearest neighbours. Scikit-learn's KNeighborsClassifier was used to implement the KNN model. The baseline model simply predicts the working day based on the majority class in the training data. In most training sets, "working day = yes" was the majority class in this scenario, so the baseline model would often classify the test set as "working day = yes." There is no class imbalance issue (68% positive class).

The three models were trained using two-level cross-validation and their performances were evaluated based on their accuracy and confusion matrices. $K_1 = K_2 = 10$ was used. The same D_{par}^{i}, D_{test}^{i} splits were used to retrain the 3 methods.

	KNN		**Logistic Regression**		**Baseline**
i	k_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	4	0.095	1e-0.758	0.108	0.378
2	5	0.082	1e-0.899	0.027	0.288
3	5	0.151	1e-0.899	0.041	0.301
4	5	0.178	1e-0.899	0.055	0.219
5	5	0.123	1e-0.899	0.068	0.315
6	5	0.137	1e-0.899	0.055	0.260

7	5	0.178	1e-0.899	0.068	0.315
8	5	0.137	1e-0.899	0.123	0.356
9	5	0.096	1e-0.899	0.082	0.438
10	5	0.123	1e-0.899	0.068	0.288

Table 3. Two-level cross-validation table used to compare the three models for classification, rounded to 3dp.

From table 3, it can be seen that the regularised logistic regression and KNN models consistently had lower test errors than the baseline model. And the regularised logistic regression model had lower test errors than the KNN model.

Statistical Evaluation

Statistical evaluation was carried out following setup I using McNemar's tests at α=0.05 significance level. Thus, it should be noted that the conclusions established in the following analysis are dependent on this specific bike sharing dataset.

Model A	KNN	KNN	Regularised Logistic Regression
Model B	Regularised Logistic Regression	Baseline	Baseline
Comparison Matrix	[616 20] [64 31]	[475 161] [25 70]	[482 198] [18 33]
CI	(-0.0843, -0.0360)	(0.152, 0.220)	(0.211, 0.281)
p-value	1.59e-06	1.57e-25	1.65e-39
θ^	-0.0602	0.186	0.246
Performance Conclusion	$M_{\text{RegLogRegression}} > M_{\text{KNN}}$	$M_{\text{KNN}} > M_{\text{baseline}}$	$M_{\text{RegLogRegression}} > M_{\text{baseline}}$

Table 4. Relevant statistics obtained from comparing the 3 methods pairwise, rounded to 3sf.

The regularised logistic regression, KNN, and baseline models had an accuracy of 680 (93.0%), 636 (87.0%), and 500 (68.4%) respectively. Since all 3 p-values are well below the 0.05 significance level, the results observed are unlikely due to chance, so it is unlikely that the three models have the same performance. From the confidence intervals, it can be concluded that performance of $M_{\text{RegLogRegression}} > M_{\text{KNN}} > M_{\text{baseline}}$ when trained on this specific bike sharing dataset at the significance level of 0.05.

According to these findings, the regularized logistic regression model is the best model for the classification problem of this dataset. However, the performance difference between the KNN and regularised logistic regression is not drastic, and the choice of model may be influenced by factors such as interpretability, processing resources, and other problem-specific needs of the bike rental operator.

In the provided code, a logistic regression model is trained using the L2 regularization penalty.. The logistic regression model makes a prediction based on the logistic function, which outputs a probability value between 0 and 1 for each observation. If the probability value is above a threshold (default 0.5 in scikit-learn), the observation is classified as belonging to the positive class (working day), otherwise it is classified as belonging to the negative class (non-working day).

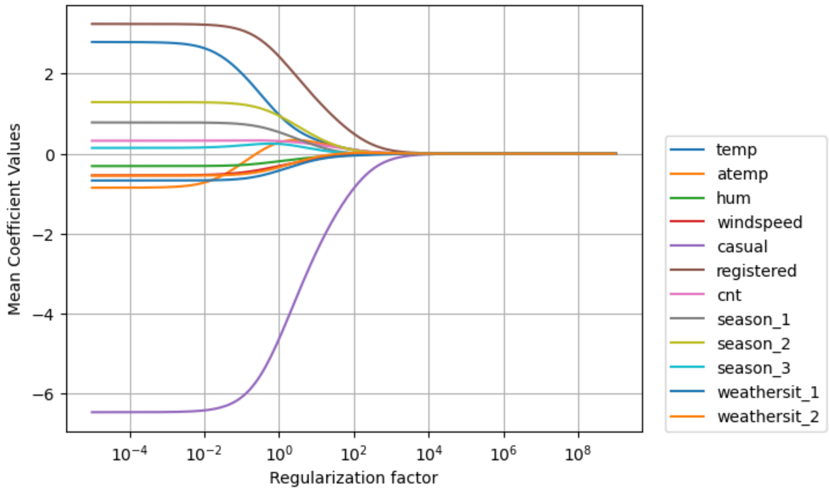


Figure 5. Graph of mean coefficient values against regularisation factor for logistic regression.

From figure 5, casual users, registered users and temperature are the stronger indicators of working days. High values of registered users and temperature and low values of casual users is more likely to result in a working day prediction. However, it is unexpected to note that the atemp and temp factors have opposite weights as these factors are closely related in reality. Temp is the only relevant factor in both the regression and classification problem. Otherwise, there is no other clear similarity of the features deemed relevant between the regression and classification problem.

<u>Discussion:</u>

We learnt about numerous classification approaches, such as KNN and logistic regression, in the report's classification section. We also learned about regularization and how to utilize it to reduce model complexity. To assess the performance of different models, we employed two-level cross-validation and evaluated them using statistical metrics such as accuracy, confusion matrix, p-values, and confidence intervals.

Overall, we discovered that selecting the suitable model and regularization parameter can have a considerable influence on the model's performance. We also discovered that assessing and evaluating the outcomes of a regression or classification model is an iterative process that requires careful consideration of statistical metrics as well as the context of the situation at hand.

Furthermore, we discovered that several characteristics were considered meaningful for both regression and classification models, implying that they may have a major influence on the problem's conclusion. Overall, we learnt strategies and methodologies that may be used for a wide range of data analysis and modelling challenges.

Previous Studies:

There were no relevant previous studies found in relation to the regression problem for estimating total daily count of bike users.

1. "Machine Learning Applied to Bike Sharing Classification Using the Washington D.C. Bike Sharing Dataset"

Authors: David Bertolino, Juan Gómez Romero, Carlos Rivas Bolaños, and Diego Zapata-Rivera

This paper gives a classification analysis using the same bike sharing dataset used in this report.. However, the target variable was a different variable from our analysis as they attempted to forecast whether a bike sharing journey would be short or lengthy. The authors employ a variety of classification methods, including decision trees, random forests, support vector machines, and neural networks. The research also uses cross-validation and statistical analysis to evaluate the algorithms' performance. The results demonstrate that the random forest method performs best, with an accuracy of 78.8%, and may be used to forecast the length of bike sharing trips in Washington, D.C.

Exam Problems:

Question 1. Spring 2019 question 13. D is the correct answer

Base on description, the observations are plotted horizontally based on their predicted value $\widehat{yi}$ and marker/color indicate the class membership. Looking at the ROC curve in figure 1, we see that it is formed in black and red markers moving upward and to the right. This indicates the classifier is performing well and has high true positive rate with low false positive. **D** has a marker gradually moving upward to right with a mix of black and red markers which is consistent to the ROC curve in fig 1, this leads to prediction D corresponds to the ROC.

Question 2. Spring 2019 question 15 : B The impurity gain of the split x7 = 2 is Δ ≈ 0.0178

Class Error = 1 - max p(c|v)
For X7 = 0, Parent node impurity : I(r)= 1 - max(33/120, 28/120, 30/120, 29/120) = 1 - max(0.275, 0.233, 0.25, 0.242) = 0.725
For X7 = 1
I(parent) = 1 - max(4/11, 2/11, 3/11, 2/11) = 1 - max(0.364, 0.182, 0.273, 0.182) = 0.636
For X7 = 2, we have:
I(parent) = 1 - max(0/4, 1/4, 0/4, 0/4) = 1 - max(0, 0.25, 0, 0) = 0.75

Left Child node impurity: I(left)= 1 - max (0/4, 1/4, 0/4, 0/4) = 1 - max (0, 0.25 ,0, 0) = 0.75
Right child node impurity: I(right) = 1 - max(33/116, 27/116, 30/116, 29/116) = 1 - max(0.284, 0.233, 0.259, 0.25) = 0.716
Where left child and right child impurity are proportions of observation in the left and right child nodes respectively. Then for X7 = 2 we have:

$$\Delta = I(r) - \sum_{k=1}^{K} \frac{N(v_k)}{N(r)} I(v_k)$$

Δ = 0.75 - (4/120 * 0.75 + 116/120 * 0.716) = ≈ 0.0178. This lead B is correct answer

Question 3. Spring 2019 question 18: A is the correct : 124 parameters

The number of parameters in a neural network is calculated as the sum of the number of each parameter in each layer. We have a single hidden layer with 10 units and a softmax output layer with 4 units. For the hidden layer, each unit has 7 inputs and a bias term,given a total of 8 parameter pre units. The hidden layer has a total 10 X 8 = 80. For the output layer,each unit has 10 inputs and bias terms, giving a total of 11. Therefore the output layer has a total of 4 X 11 = 44.The total number of parameters in Neural Network is 80 + 44 = 124

Question 4. Spring 2019 question 20: D is the correct answer

When we look at figure4 in question, until we reach the correct variation by testing all possibilities. For first PCA, which it divided the plot b_1 at -0.76 and classify all areas which are larger than -0.16 as in congestion level 4 according to PCA 1 . At cut D which it divide the area between -0.76 < b_1 < -0.16 with b_2 >= 0.01. This can classify it as congestion level3 or one part of congestion level2, while the rest of two get classified after A with B as b_2 >= 0.01. This leads to the congestion level one is defined, so the correct answer is D A: $b_1 \geq -0.76$, B: $b_2 \geq 0.03$, C: $b_1 \geq -0.16$, D: $b_2 \geq 0.01$

Question 5. Spring 2019 question 22: C is the correct : 3570.0

We know form problem statement the inner cross validation K2= 4 folds, and test values for λ: { 0.01, 0.06, 0.32, 1.78, 10} nh : {1,2,3,4,5}. For each Outer fold, we will train and test 25 NN models and 5 logistic regression models.first we calculate the time taken to train and test single NN model for one inner fold and one outer fold: training time : 20 ms, testing time: 5 m, so

One inner fold and one outer fold , the total time to train and test all 25 NN model is: 25 * (20+5) = 625 ms. similarly, for one inner fold and one outer fold, the total time taken to train and test all 5 logistic regression model is : 5*(8 +1) = 45 ms. Since we have 5 outer folds, the total time taken to train and test all models for all outer folds is:(625 + 45) *5 = 3350 ms. As the table shows only the optimal parameters and corresponding test errors for each outer fold, we assume the time taken to compose the table is negligible compared to time to train and test the models , therefore the answer C is closed option and the correct.

Question 6. Spring 2019 question 26: B is the correct answer

We have to calculate the following: $\widehat{Y}$ k = $\left[1\ b_1\ b_2\right]$^T W_k , for k =1, ... , 3 for each observations by using softmax transformation equation

$$P(y=k|\hat{\boldsymbol{y}}) = \begin{cases} \frac{e^{\hat{y}_k}}{1+\sum_{k'=1}^{3} e^{\hat{y}_{k'}}}. & \text{if } k \leq 3 \\ \frac{1}{1+\sum_{k'=1}^{3} e^{\hat{y}_{k'}}} & \text{if } k = 4 \end{cases}$$

Then we need to calculate and sum *all* e^{yk} in denominators, and add a constant 1 to it. In the question, they ask about which observation will be assigned to class y =4, so each observation of probability for 4 should be calculated. The correct option appears after we did calculation is **B**