Упражнения за курса Математически основи на машинното самообучение

Абстракт

Упражнения за курса Математически основи на машинното самообучение

Съдържание

1 Уводни бележки. Базов Python. Линейна регресия.

1.1 Уводни бележки

Мотивация - Курсът има широко приложение в индустрията.
Литература:

1. Introduction to statistical learning [2].

2. Elements of Statistical learning [1].

Какво ще съдържа курса? Приблизително съдържанието на тези записки.

Исторически бележки - Vapnik theory.(може да се добавят няколко изречения)

Среда за писане на код - VsCode и/или Anaconda + Spyder, Jupyter Notebook. Забележка: преди 2 години Anaconda стана платена при enterprise, имайте го предвид.
Тук ще изпозлваме Python 3.10.11, последната версия с prebuild инсталационен пакет от python.org, от версиите 3.10.11. За 3.11 и наскоро излязлата 3.12 нямаме още всички пакети. Ако ще изпозлваме виртуална среда:

```
$pip install virtualenv
```

(Да се добавят git clone и прочие... команди и и описания)

Отваряме с VSCode папката с кода на упражненията и пишем:

```
$pip install −r requirements.txt
```

За инсталиране на пакетите, които се използват в курса.
Линк към страницата с Lab упражненията от книгата:
https://intro-stat-learning.github.io/ISLP/installation.html

1.2 Базов Python

Най-базови неща в Python - python list, set, import packages, най-базови фунцкии в numpy, събиране на вектори и умножение на вектори с число, скаларно произведение, матрици. pandas пакета и пример за графика с matplotlib.pyplot.

Lab 2.3 Introduction to Python от Python книгата. Линк:

https://intro-stat-learning.github.io/ISLP/labs/Ch02-statlearn-lab.html#.

1.3 Линейна регресия

Данни, разглеждане - упражнение 2.8, 2.9 или 2.10 от Applied.
(!) Добре е или да се разгледат функциите load_data() от ISLP пакета или да се изнесе частта от кода за конкретния dataset при упражненията.

Преговаряне на метод на най-малките квадрати:
Нека имаме множество от точки $(x_1, y_1), (x_n, y_n), \dots (x_n, y_n)$. Права от вида $\hat{y} = \hat{\beta}_0 + \sum \hat{\beta}_i x_i$, която минимизира $\sum (y - \hat{y})^2$ може да се намери с метод на най-малките квадрати.
С какво е полезна линейната регресия? Може да се използва за "старт".
Плюсове и минуси - прост модел, лесно се смята и разбира, лесно се обяснява, недоба в нелинейни задачи.

Да се добави Multiple regression и значение на променливите, примерът на Hastie and Tibshirani с успоредни прави.

Следва Lab 3.6

1.4 Упражнения

В зависимост от отделеното време на въведение в Python, 2 от 3 по-долу може би ще е добре

Упажнение 1.4.1 (2.8) *This exercise relates to the College data set, which can be found in the file College.csv on the book website. It contains a number of variables for 777 different universities and colleges in the US. The variables are*

- **Private** *: Public/private indicator*
- *Apps : Number of applications received*
- *Accept : Number of applicants accepted*
- *Enroll : Number of new students enrolled*
- *Top10perc : New students from top 10 % of high school class*
- *Top25perc : New students from top 25 % of high school class*
- *F.Undergrad : Number of full-time undergraduates*
- *P.Undergrad : Number of part-time undergraduates*

- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- PhD : Percent of faculty with Ph.D.s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni wh
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into Python, it can be viewed in Excel or a text editor.

(a) Use the pd.read_csv() function to read the data into Python. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

(б) Look at the data used in the notebook by creating and running a new cell with just the code college in it. You should notice that the first column is just the name of each university in a column named something like Unnamed: 0. We don't really want pandas to treat this as data. However, it may be handy to have these names for later. Try the following commands and similarly look at the resulting data frames:

```
college2 = pd.read_csv('College.csv', index_col=0)
college3 = college.rename({'Unnamed: 0': 'College'},
                          axis=1)
college3 = college3.set_index('College')
```

This has used the first column in the file as an index for the data frame. This means that pandas has given each row a name corresponding to the appropriate university. Now you should see that the first data column is Private. Note that the names of the colleges appear on the left of the table. We also introduced a new python object above: a dictionary, which is specified by dictionary (key, value) pairs. Keep your modified version of the data with the following:

```
  college = college3
```

(в) Use the describe() method of to produce a numerical summary of the variables in the data set.

(г) Use the pd.plotting.scatter_matrix() function to produce a scatterplot matrix of the first columns [Top10perc, Apps, Enroll]. Recall that you can reference a list C of columns of a data frame A using A[C].

(д) *Use the boxplot() method of college to produce side-by-side boxplots of Outstate versus Private.*

(е) *Create a new qualitative variable, called Elite, by binning the Top10perc variable into two groups based on whether or not the proportion of students coming from the top 10school classes exceeds 50*

```
college[’Elite’] = pd.cut(college[’Top10perc’],
[0,0.5,1],
labels=[’No’, ’Yes’])
```

(ж) *Use the value_counts() method of college[’Elite’] to see how many elite universities there are. Finally, use the boxplot() method again to produce side-by-side boxplots of Outstate versus Elite.*

(з) *Use the plot.hist() method of college to produce some his- tograms with differing numbers of bins for a few of the quanti- tative variables. The command plt.subplots(2, 2) may be use- ful: it will divide the plot window into four regions so that four plots can be made simultaneously. By changing the arguments you can divide the screen up in other combinations.*

(и) *Continue exploring the data, and provide a brief summary of what you discover.*

Упажнение 1.4.2 (2.9) *This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.*

(а) *Which of the predictors are quantitative, and which are quali- tative?*

(б) *What is the range of each quantitative predictor? You can an- swer this using the min() and max() methods in numpy.*

(в) *What is the mean and standard deviation of each quantitative predictor?*

(г) *Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?*

(д) *Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.*

(е) *Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.*

Упажнение 1.4.3 (2.10) *This exercise involves the Boston housing data set.*

(а) *To begin, load in the Boston data set, which is part of the ISLP library.*

(б) *How many rows are in this data set? How many columns? What do the rows and columns represent?*

(в) *Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.*

(г) *Are any of the predictors associated with per capita crime rate? If so, explain the relationship.*

(д) *Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.*

(е) *How many of the suburbs in this data set bound the Charles river?*

(ж) *What is the median pupil-teacher ratio among the towns in this data set?*

(з) *Which suburb of Boston has lowest median value of owner- occupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.*

(и) *In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.*

2 Логистична регресия. Linear Discriminant analysis.

2.1 Кратка информация

Мотивация - логистичната регресия се използва за Credit Score модели. Моделът има вида:
$\mathbb{E}(Y|X_1,\dots,X_p) = Prob(Y=1|X_1,\dots,X_p) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$
$\mathbb{E}(Y|X_1,\dots,X_p) = Prob(Y=0|X_1,\dots,X_p) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$

Идея за други класификационни модели - Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes. Naive Bayes - ....
Linear Discriminant analysis
Ако класовете са с многомерно гаусово разпределение, или

$$f_k(x) = \frac{1}{(2\pi)^p/2\,|\Sigma_k|}e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

в специалния случай, където ковариационните матрици са равни, $\Sigma_k = \Sigma \forall k$, имаме Линеен дискриминантен анализ.

Теорема на Bayes с доказателство(припомняне):

Theorem 1 (Bayes) $P(A|B) = \frac{P(B|A)P(B)}{P(A)}$.

доказателство: По дефиниция $P(A|B) = P(A\cap B)P(B)$ и $P(B|A) = P(B\cap A)P(A)$. Получаваме $P(A\cap B) = \frac{P(A|B)}{P(B)} = \frac{P(B|A)}{P(A)}$, откъдето следва равенството. ■

2.2 Упражнения

Conceptual 4.8 section, ex 1, 2, 12

Упажнение 2.2.1 (Sec 4.8, Conceptual ex. 1) *Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit represen- tation for the logistic regression model are equivalent.*

Упажнение 2.2.2 (Sec 4.8, Conceptual ex. 2) *A*

Упажнение 2.2.3 (Sec 4.8, Conceptual ex. 12) *Suppose that you wish to classify an observation $x \in R$ into apples and oranges. You fit a logistic regression model and find that*

Pr(Y = orange|X = x) = . Your friend fits a logistic regression model to the same data using the softmax formulation in (4.13), and finds that

$$Pr(Y = orange|X = x) = \frac{exp(\hat{a}_{orange} + \hat{a}_{orangex})}{den}$$

$Pr(Y = orange/X = x) =$ (a) What is the log odds of orange versus apple in your model? (b) What is the log odds of orange versus apple in your friend's model? (c) Suppose that in your model, β_0 and β_1. What are the coefficient estimates in your friend's model? Be as specific as possible. (d) Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient 0.6. What are the coefficient estimates in your model? (e) Finally, suppose you apply both models from (d) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

3 Feature Selection. Model Selection - AIC, BIC. Bias-Variance tradeoff. K-fold Cross Validation.

Weight of evidence, information value - използват се в индустрията, например при разработка на credit score модел, логистична регресия.
В книгата има 4 подхода при избор на оптимален модел:
C_p статистика: $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$ за squared error loss

- $AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$
- $BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$
- $Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$

(да се добави значението на параметрите по-късно)
(код за пресмятане с пример)

lab 6 за model selection
Forward, Backward selection example
Следните алгоритми са взети от книгата:
Алгоритъм за best subset selection:

1. Let M0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For k = 1, 2, . . . p:
   (a) Fit all (?) p choose k models that contain exactly k predictors.
   (б) Pick the best among these (?) p choose k models, and call it Mk. Here best is defined as having the smallest RSS, or equivalently largest R2.
3. Select a single best model from among M0, . . . , Mp using using the prediction error on a validation set, Cp (AIC), BIC, or adjusted R2. Or use the cross-validation method

Алгоритъм за Forward selection:

1. Let M0 denote the null model, which contains no predictors.
2. For k = 1, 2, . . . p:
   (a) Consider all $p-k$ models that augment the predictors in Mk with one additional predictor.
   (б) Choose the best among these$p-k$ models, and call it Mk+1. Here best is defined as having smallest RSS or highest R2.
3. Select a single best model from among M0, . . . , Mp using the pre- diction error on a validation set, Cp (AIC), BIC, or adjusted R2. Or use the cross-validation method.

Credit data

5.3 Lab python book: Cross-Validation and the Bootstrap

4 Generalized linear models. Lasso, Ridge. Generalized additive models.

4.1 Generalized linear models

Lasso and Ridge from the book.
Lasso formula, Ridge formula, explanation
Въпрос: Как Ridge и Lasso fit-ват в общите линейни модели или общите адитивни модели?

Да припомним няколко регресии, които сме виждали:

- линейна - $\mathbb{E}(Y|X_1,\dots,X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$
- логистична - $\mathbb{E}(Y|X_1,\dots,X_p) = \frac{e^{\beta_0+\beta_1 X_1+\dots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\dots+\beta_p X_p}}$
- Поансонова - $\mathbb{E}(Y|X_1,\dots,X_p) = e^{\beta_0+\beta_1 X_1+\dots+\beta_p X_p}$

Общият линеен модел изглежда по следния начин(където η е "свързваща фунцкия"):

$$\eta(\mathbb{E}(Y|X_1,\dots,X_p)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

lab 6 за ridge lasso
Change λ for Lasso example.

Подробно разписване на оптимизационната задача. Идея за множество от бъдещите задачи се разглеждат като оптимизационни.

5 Нелинейни модели. Piecewise polynomials. Сплайни(Splines). Общи адитивни модели

Lab 7.8 python
ex 7.9 6,10,11

5.1 Общи адитивни модели(Generalized additive models)

Това е 7.7 от Python книгата.

Общ адитивен модел се за регресия се задава с уравненията(където заменяме x_i с $f(x_i)$ от модела на линейна регресия):

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i = \beta_0 + f_1(x_{i1}) + \cdots + f_n(x_{ip}) + e_i$$

за $i = 0, 1 \ldots n$. Общ адитивен модел се за класификация се задава с формулите:

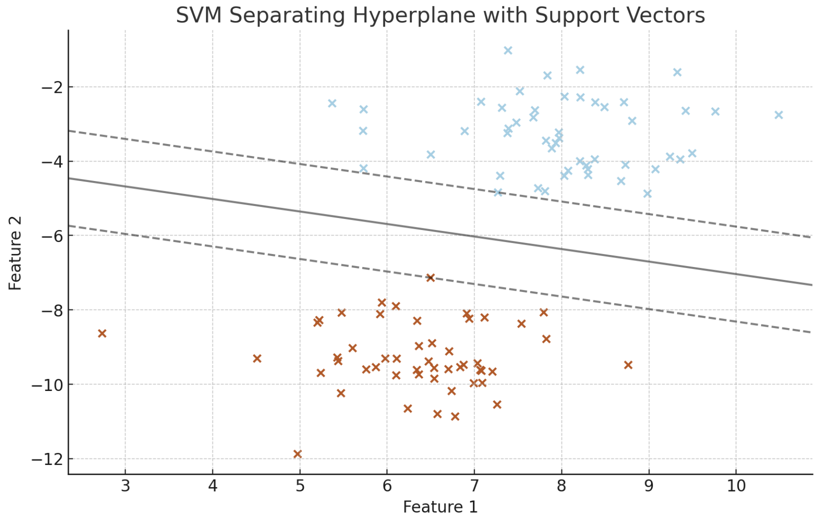
$$log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 f_1(X_1) + \beta_2 f_2(X_2) + \cdots + \beta_p f_p(X_p),$$

където отново "сме заменили X-овете от логистичната регресия с $f(X)$-ове":

$$log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

6 Decision trees. Random fores. Bagging, boosting.

7 Support Vector Machines. Метод на опорните вектори.



Разделяща хиперравнина.

Recall: Хиперравнина се задава с $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = 0$. Нормален вектор на тази хиперравнина е $(\beta_1, \dots \beta_p)$ и той задава положителната посока. Уравнението може да се запише и като $\beta_0 + (\boldsymbol{\beta}, \mathbf{X}) = 0$, където $\boldsymbol{\beta} = (\beta_1, \dots \beta_p)$ и $\mathbf{X} = (X_1, \dots, X_p)$. Една хиперравнина разделя пространството на две части, една с $\beta_0 + (\boldsymbol{\beta}, \mathbf{X}) > 0$, и една с $\beta_0 + (\boldsymbol{\beta}, \mathbf{X}) < 0$. Ако имаме точки в пространството, които можем да разделим с хиперравнина, и дадем labels на тези точки $y_i = 1$ и $y_i = -1$, то

$$y_i(\beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip}) > 0.$$

Оптимизационна задача:

$$\text{да се максимизира } M \tag{1}$$

$$\text{така че } \sum_{j=1}^{p} \beta_j^2 = 1 \tag{2}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip}) \geq M \text{ за } i = 1, \dots, n \tag{3}$$

Втора оптимизационна задача(Support Vector Classifier):

$$\text{да се максимизира } M \tag{4}$$

$$\text{така че } \sum_{j=1}^{p} \beta_j^2 = 1 \tag{5}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip}) \geq M(1 - \epsilon_i) \text{ за } i = 1, \dots, n \tag{6}$$

$$\text{и } \epsilon_i \geq 0, \sum_i e_i = C \tag{7}$$

Support Vector Machines:
Решението на втора оптимизационна задача може да се запише във вида

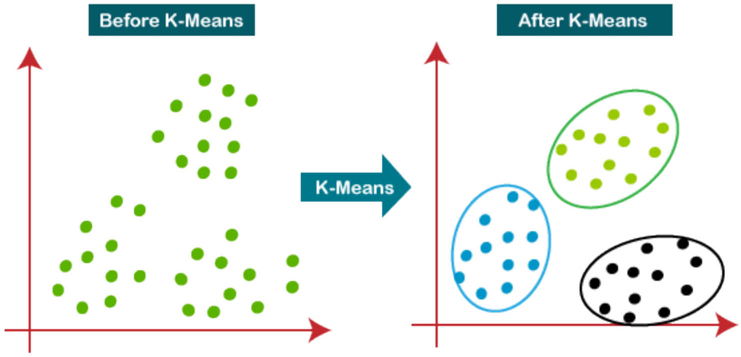
$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle,$$

и тогава на нас са ни достатъчни скаларните произведения $\langle x_i, x_j' \rangle$ за всички $\binom{n}{2}$ двойки x_i, x_j', $i, j = 1, \dots, n$, за да пресметнем модела. Вместо $\langle x_i, x_j' \rangle$, може да вземем произволна kernel фунцкия $K(x_i, x_j')$. Примерът със скаларното произведение е с линейна фунцкия, но може и нелинейна.

Упражнения: Conceptual 1,3

8 Clustering. K-means clustering. Higherarchical clustering, Йерархично клъстериране

8.1 Clustering. K-means clustering



Дефиниция 1 *Клъстери на множеството* $\{1, \dots, n\}$ *наричаме подмножества* $C_1 \cup \dots \cup C_K$*, за които е изпълнено:*

1. $C_1 \cup \dots \cup C_K = \{1, \dots, n\}$,
2. $C_i \cap C_j = \emptyset$ *за всички* $i \neq j$.

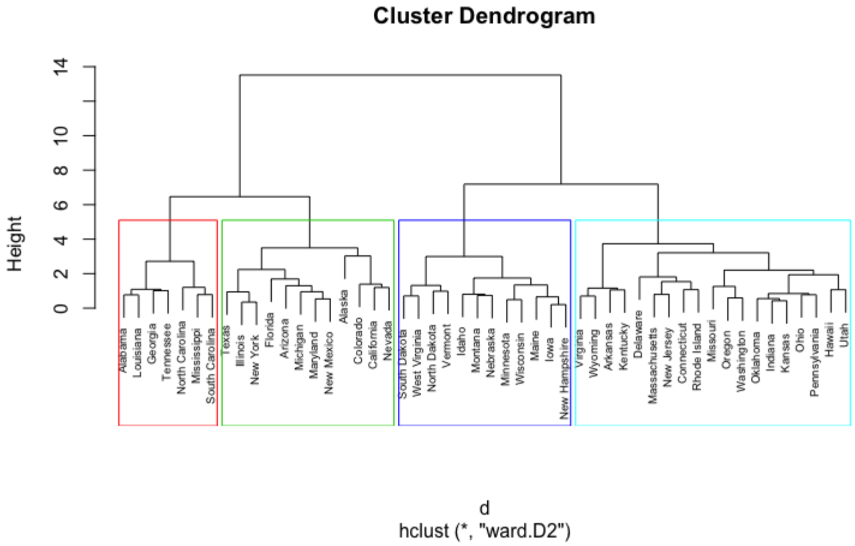
Дефинираме оптимизационна задача:

$$\text{minimize}\Big\{\sum_{j=i}^{K} W(C_i)\Big\} \tag{8}$$

$$W(C_i) = \frac{1}{C_i} \sum_{j,l \in C_i} \sum_{s=1}^{p} (x_{js} - x_{ls})^2 \tag{9}$$

Алгоритъм за K-means:
Algo

8.2 Higherarchical clustering



Упражнения:
K means ex 1 and 3, 10
Chapter 12, ex 13, hierarchical

[1]

Литература

[1] Trevor Hastie, Robert Tibshirani и Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[2] Gareth James и др. *An Introduction to Statistical Learning with Applications in Python*. Springer Texts in Statistics. Cham: Springer, 2023. ISBN: 978-3-031-38746-3. DOI: `10.1007/978-3-031-38747-0`. URL: `https://link.springer.com/book/10.1007/978-3-031-38747-0`.