دانشگاه شهید بهشتی
دانشکده علوم ریاضی
گروه علوم کامپیوتر

تمرین های سری دوم
درس داده‌کاوی

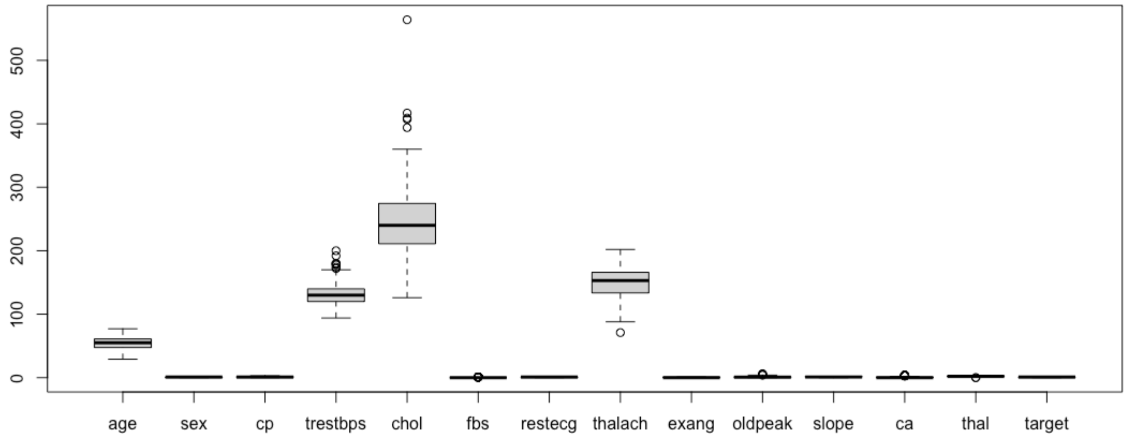
والا خسروی
۹۹۴۲۲۰۶۸

جناب آقای دکتر فراهانی
جناب آقای دکتر خردپیشه

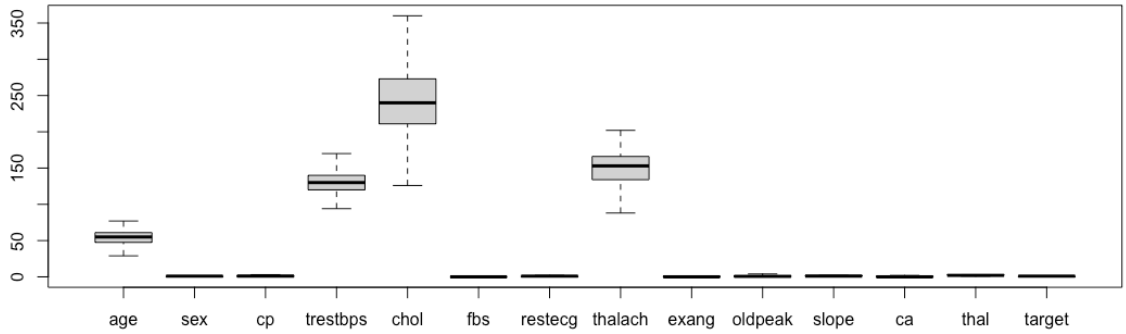
بهار ۱۴۰۰

سوال ۱

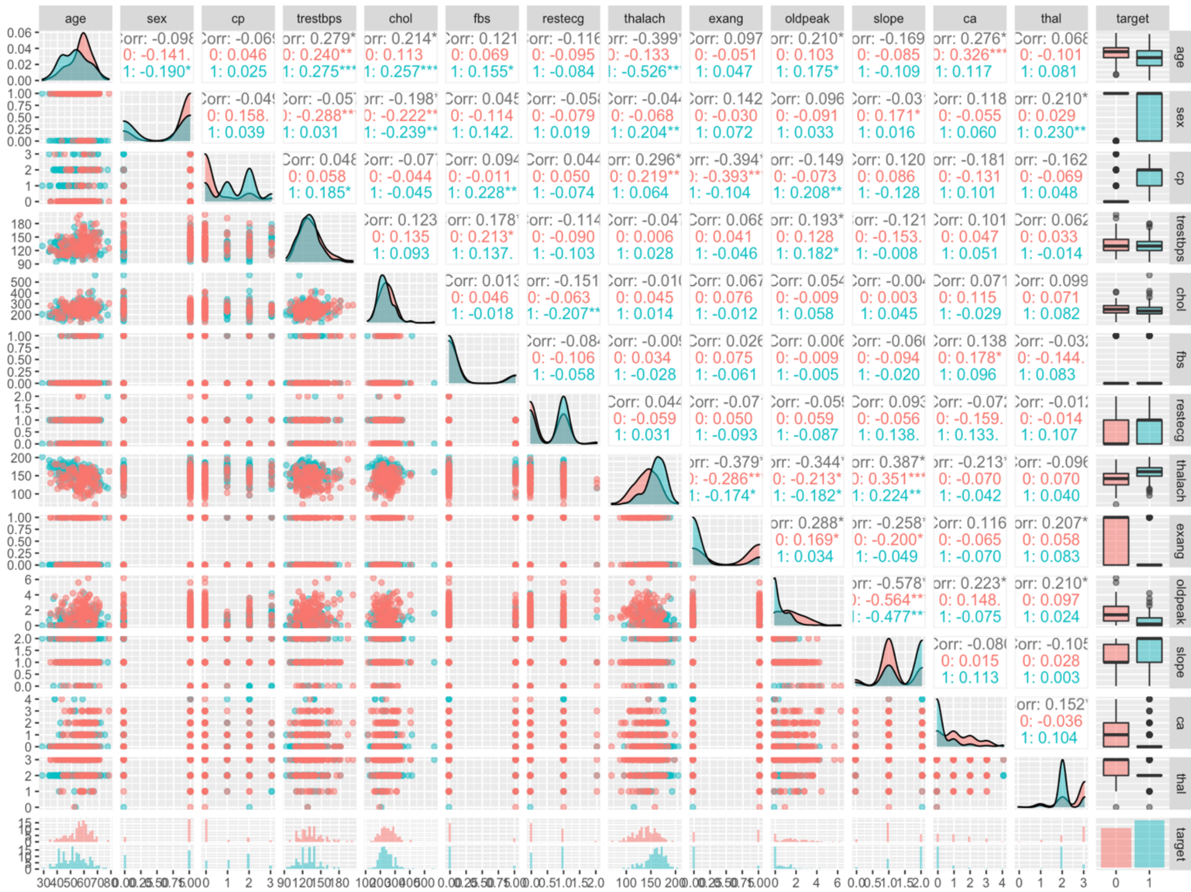
- بله، با توجه به شکل زیر داده پرت وجود دارد. نقاط خارج از حدود نمودارهای جعبه‌ای داده پرت هستند.



بعد از حذف داده‌ها به شکل زیر تبدیل می‌شوند.



- با توجه به نمودار زیر داده‌ها در بعضی ستون‌ها متوازن هستند ولی در بعضی ستون‌ها خیر (مثل ستون sex, cp exang, thal, ca)



برای حل مشکل کلاس های نامتوازن در الگوریتم‌های پیش‌بینی رویکردهای مختلفی مواجهه با داده‌های نامتوازن وجود دارند مانند:

الف) رویکرد در سطح داده: تکنیک‌های Resampling

Informed Over Sampling: Synthetic Minority Over Sampling Technique

Cluster-Based Over Sampling

Random Under Sampling

Random Over Sampling

Modified synthetic minority oversampling technique (MSMOTE)

ب) تکنیک های الگوریتمی تجمعی (Algorithmic Ensemble Techniques)

Bagging Based Boosting-Based
Adaptive Boosting- Ada Boost Gradient Tree Boosting
XG Boost

سوال ۲

در زبان r با دستورهای زیر انجام می‌شود

```r
train_ind <- sample(nrow(df), 242, replace = FALSE, prob = NULL)
train <- df[train_ind, ]
test <- df[-train_ind, ]
```

سوال ۳

قضیه بیز از روشی برای دسته‌بندی پدیده‌ها بر پایه احتمال استفاده میکند و احتمال رخ احتمال رخداد پیشامد A به شرط B برابر است با احتمال رخداد پیشامد B به شرط A ضرب در احتمال رخداد پیشامد A تقسیم بر احتمال رخداد پیشامد B.

دسته بند بیز ساده گاوسی

اگر مشاهدات و داده‌ها از نوع پیوسته باشند، از مدل احتمالی با توزیع گاوسی یا نرمال برای متغیرهای مربوط به شواهد می‌توانید استفاده کنید. در این حالت هر دسته یا گروه دارای توزیع گاوسی است. به این ترتیب اگر k دسته یا کلاس داشته باشیم می‌توانیم برای هر دسته میانگین و واریانس را محاسبه کرده و پارامترهای توزیع نرمال را برای آن‌ها برآورد کنیم.

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

دسته بند بیز ساده چندجمله‌ای

بیز ساده چندجمله‌ای، به عنوان یک دسته‌بند متنی بسیار به کار می‌آید. در این حالت برحسب مدل احتمالی یا توزیع چند جمله‌ای، برداری از n ویژگی برای یک مشاهده به صورت $X = (x_1, \dots, x_n)$ با احتمالات $(p_1, \dots, p_n)$ در نظر گرفته می‌شود. مشخص است که در این حالت بردار X بیانگر تعداد مشاهداتی است که ویژگی خاصی را دارا هستند. به این ترتیب تابع درستنمایی در چنین مدلی به شکل زیر نوشته می‌شود.

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

دسته بند بیز ساده برنولی

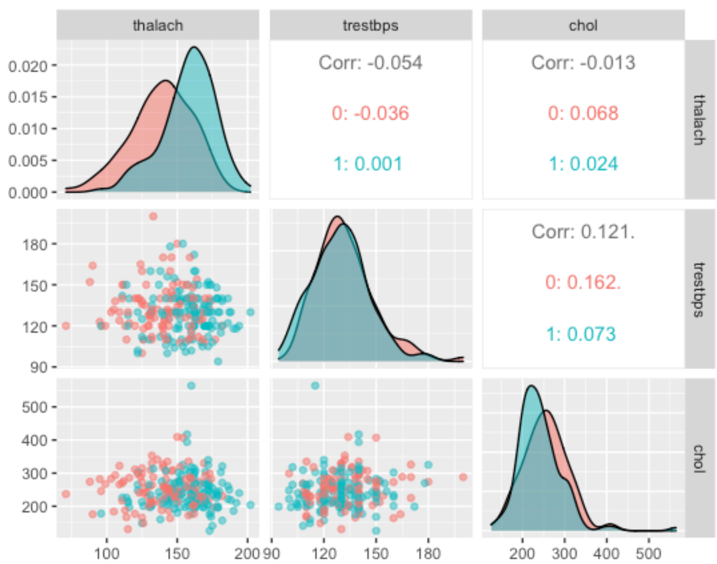
در این قسمت به بررسی توزیع برنولی و دسته‌بندی بیز خواهیم پرداخت. به شکلی این نوع از دسته‌بند بیز بیشترین کاربرد را در دسته‌بندی متن‌های کوتاه داشته، به همین دلیل محبوبیت بیشتری نیز دارد. در این مدل در حالت چند متغیره، فرض بر این است که وجود یا ناموجود بودن یک ویژگی در نظر گرفته شود. برای مثال با توجه به یک لغتنامه مربوط به اصطلاحات ورزشی، متن دلخواهی مورد تجزیه و تحلیل قرار می‌گیرد و بررسی می‌شود که آیا کلمات مربوط به لغتنامه ورزشی در متن وجود دارند یا خیر. به این ترتیب مدل تابع درستنمایی متن براساس کلاس های مختلف C_k به شکل زیر نوشته می‌شود.

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)}$$

سوال ۴

پیاده سازی در فایل script.r با نام gaussian_naive_bayes پیاده سازی شده است.

این شکل در سوال‌های پیش رو قابل استفاده است. (به خصوص مسائل svm)

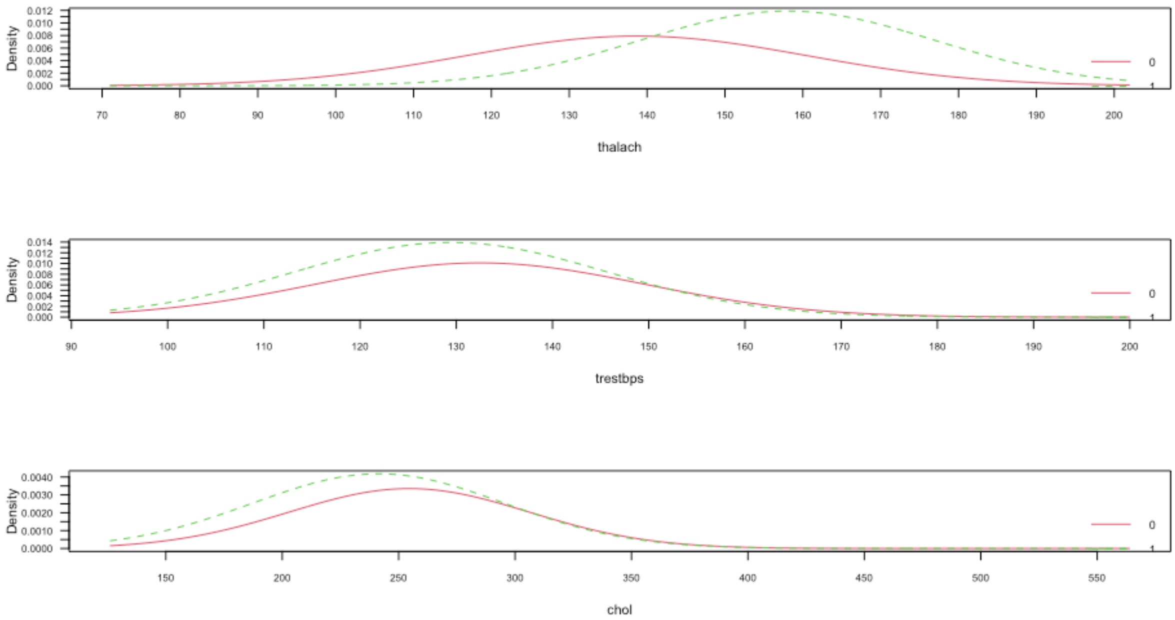


سوال ۶

مدل به شکل زیر تعریف شده است.

```
======================================= Naive Bayes ========================================

 Call:
naive_bayes.formula(formula = target ~ thalach + trestbps + chol,
    data = train)

--------------------------------------------------------------------------------------------

Laplace smoothing: 0

--------------------------------------------------------------------------------------------

 A priori probabilities:

        0         1
0.4338843 0.5661157

--------------------------------------------------------------------------------------------

 Tables:

--------------------------------------------------------------------------------------------
 ::: thalach (Gaussian)
--------------------------------------------------------------------------------------------

thalach         0         1
   mean 138.65714 158.02920
   sd    21.89039  19.00501

--------------------------------------------------------------------------------------------

--------------------------------------------------------------------------------------------
 ::: trestbps (Gaussian)
--------------------------------------------------------------------------------------------

trestbps         0         1
    mean 132.46667 129.41606
    sd    17.12402  16.22074

--------------------------------------------------------------------------------------------
 ::: chol (Gaussian)
--------------------------------------------------------------------------------------------

chol           0         1
  mean 254.49524 241.33577
  sd    51.64619  54.02413

--------------------------------------------------------------------------------------------
```

نمودارهای احتمال آن به شکل زیر است.



و مقادیر محاسبه شده سه معیار مورد نظر در شکل زیر موجود می‌باشد.

```
> print(c('precision', precision))
[1] "precision"         "0.857142857142857"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall"            "0.571428571428571"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score"          "0.685714285714286"
```

سوال ۷

نتایج شبیه به هم بودند

سوال ۸ و ۹

مقادیر محاسبه شده سه معیار مورد نظر با کرنل **خطی** در شکل زیر موجود می‌باشد.

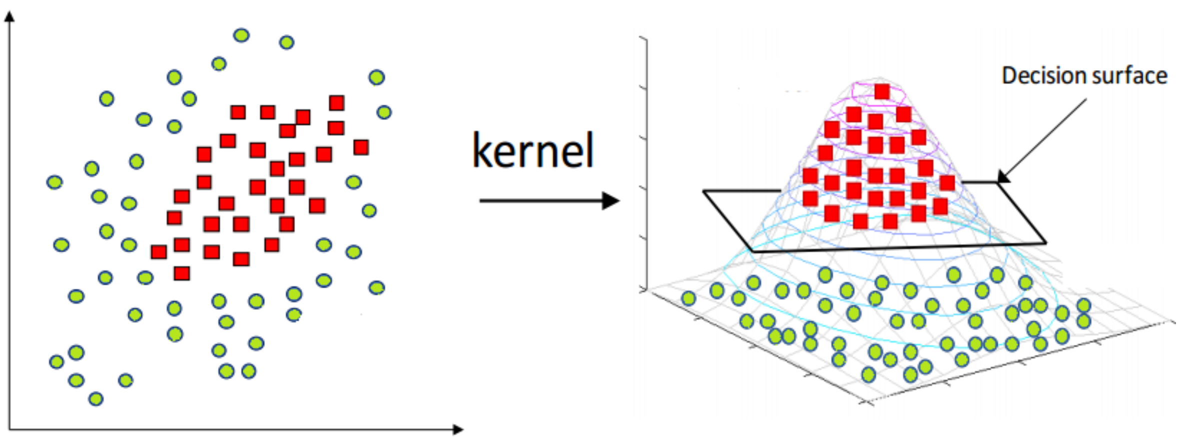
```
> print(c('precision', precision))
[1] "precision"         "0.821428571428571"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall" "0.575"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score"          "0.676470588235294"
```

مقادیر محاسبه شده سه معیار مورد نظر با کرنل **چند جمله‌ای** در شکل زیر موجود می‌باشد.

```
> print(c('precision', precision))
[1] "precision"         "0.892857142857143"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall"            "0.543478260869565"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score"          "0.675675675675676"
```

نتایج مقدار کمی تفاوت دارند

در حقیقت ، آنچه کرنل برای ما انجام می دهد ، ارائه روش کارآمدتر و کم هزینه تر برای تبدیل داده‌ها به ابعاد بالاتر است. با این گفته ، استفاده از ترفند هسته به الگوریتم SVM محدود نمی شود. هرگونه محاسبه شامل ضرب داخلی نقطه (x,y) است را می توان از ترفند کرنل استفاده کند.



سوال ۱۰

مدل به شکل زیر می‌باشد

```
> smv_model

Call:
svm(formula = target ~ ., data = train, kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  87
```

مقادیر محاسبه شده سه معیار مورد نظر با کرنل **خطی** و با در نظر گرفتن تمامی ستون‌ها در شکل زیر موجود می‌باشد.

```
> print(c('precision', precision))
[1] "precision"         "0.928571428571429"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall"            "0.702702702702703"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score" "0.8"
```

سوال ۱۱

مدل به شکل زیر می‌باشد

```
Linear Regression

303 samples
 13 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 242, 242, 242, 243, 243
Resampling results:

  RMSE       Rsquared   MAE
  0.3581397  0.4879294  0.2887452

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
> print(c('precision', precision))
[1] "precision" "0.5"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall" "0.5"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score" "0.5"
```

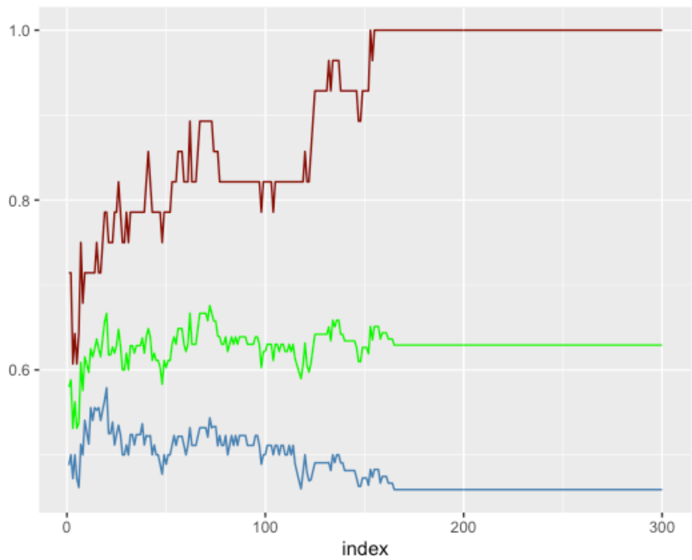
سوال ۱۲

برای K=10 نتایج به شکل زیر می‌باشد

```
> print(c('precision', precision))
[1] "precision"          "0.678571428571429"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall" "0.5"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score"           "0.575757575757576"
```

سوال ۱۳

برای حل این سوال و مقایسه مقادیر متفاوت K، از ۱ تا ۳۰۰ را محاسبه کردم و نمودار آن را رسم کردم.



قرمز تیره – percision

سبز – f1 score

آبی – recall

سوال ۱۴

```
> print(c('precision', precision))
[1] "precision"          "0.785714285714286"
> recall <- tp / (tp+fn)
> print(c('recall', recall))
[1] "recall"             "0.511627906976744"
> f1_score <- 2 * (precision * recall) / (precision + recall)
> print(c('f1_score', f1_score))
[1] "f1_score"           "0.619718309859155"
```

نتایج بهبود یافتن.

سوال ۱۵

اگر توزیع جامعه آماری نامشخص باشد و از طرفی حجم نمونه نیز کوچک باشد بطوری که نتوان از قضیه حد مرکزی برای تعیین توزیع حدی یا مجانبی جامعه آماری، استفاده کرد، از تحلیل‌های ناپارامتری استفاده می‌شود، زیرا در ایین حالت کار آمدتر از روش‌های پارامتری هستند. به این ترتیب در زمانی که توزیع جامعه مشخص نباشد و یا حجم نمونه کم باشد، روش‌های و آزمونهای ناپارامتری نسبت به روش‌ها و آزمونهای پارامتری از توان ازمون بیشتری برخوردارند و نسبت به آنها ارجح هستند.

سوال ۱۶

$$N = TN + TP + FN + FP$$

$$S = \frac{TP + FN}{N}$$

$$P = \frac{TP + FP}{N}$$

$$\mathrm{MCC} = \frac{TP/N - S \times P}{\sqrt{PS(1-S)(1-P)}}$$