دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

حل تمرین درس یادگیری ماشین

سری ۳

نام و نام خانوادگی دانشجو:

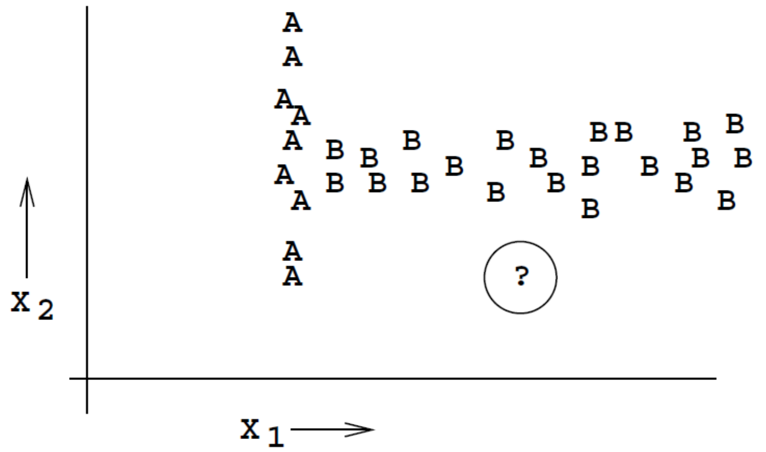
همایون حیدرزاده (۹۵۱۳۱۰۷۰)

نام استاد: دکتر ناظرفرد

آبان ۹۶

۱)

الف)



فرض میکنیم توزیع کلاس‌ها گوسی باشد بنابرین طبق شکل داریم:

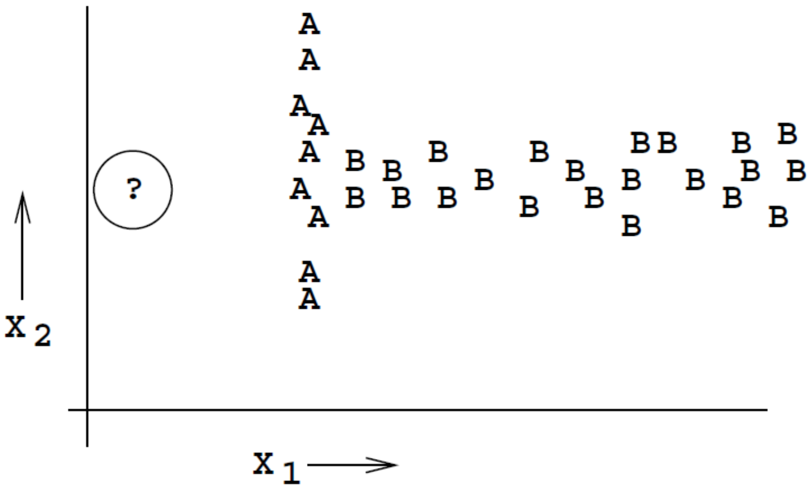
p(A)=9/30=0.3, p(B)=0.7
? -> z
p(class|z)=argmax p(z|class)p(class)
p(A|z)=p(z|A)p(A)=p(z1|A)p(z2|A)p(A)
p(B|z)=p(z|B)p(B)=p(z1|B)p(z2|B)p(A)

چون در اینجا واریانس کلاس A برای ویژگی x۱ تقریبا صفر است بنابرین احتمال p(z۱|A) بسیار کوچک می‌شود و این مقدار برای کلاس B نزدیک یک است چون نزدیک میانگین کلاس B است.

چون مقادیر احتمال برای کلاس B همگی بزرگ هستند بنابرین برچسب داده z کلاس B می‌شود. یعنی داریم:

p(B)>p(A)
p(z1|B) ~ 1.0
p(z1|A) ~ 0.0
p(z2|B) کمی بیشتر از p(z2|A)

ب)

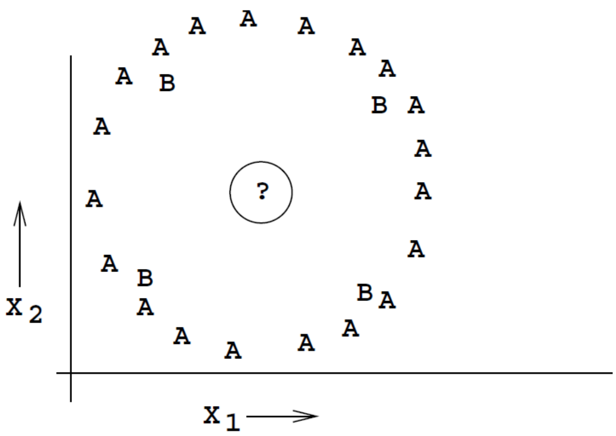


در این قسمت نیز مانند قسمت قبلی احتمال‌ها را مقایسه می‌کنیم:

p(B)>p(A)
p(z1|B) >> 0
p(z1|A) ~ 0.0
p(z2|B) ~ 1.0
p(z2|A) ~ 1.0 چون نزدیک میانگین هستند

چون در جهت X۲ داده نزدیک میانگین دو کلاس است پس احتمال آن تقریبا یک می‌شود (برای هر دو کلاس). ولی در جهت X۱ واریانس کلاس A بسیار پایین و نزدیک صفر است بنابرین احتمال p(z۱|A) باز هم نزدیک صفر است و این احتمال برای کلاس B خیلی بزرگتر است (واریانس بالا در جهت X۱). همچنین احتمال پیشین کلاس B نیز بیشتر است، بنابرین برچسب باز هم B می‌شود.

ج)



داریم:

p(A)=4/24=0.16, p(B)=0.84
? -> z
p(class|z)=argmax p(z|class)p(class)
p(A|z)=p(z|A)p(A)=p(z1|A)p(z2|A)p(A)
p(B|z)=p(z|B)p(B)=p(z1|B)p(z2|B)p(A)

اگر توزیع‌ها را گوسی در نظر بگیریم چون داده نزدیک میانگین هر دو کلاس است پس احتمال‌های زیر تقریبا ۱ می‌شوند:

p(B)>p(A)
p(z1|B) ~ 1.0
p(z1|A) ~ 1.0
p(z2|B) ~ 1.0
p(z2|A) ~ 1.0

ولی چون احتمال پیشین کلاس B خیلی بیشتر از A است پس باز هم برچسب داده جدید کلاس B خواهد بود.

---

۲)

الف)

اگر استقلالی وجود نداشته باشد، مدل ما به صورت زیر خواهد بود:

$$\begin{aligned} p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k) p(x_2, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k) p(x_2 \mid x_3, \dots, x_n, C_k) p(x_3, \dots, x_n, C_k) \\ &= \dots \\ &= p(x_1 \mid x_2, \dots, x_n, C_k) p(x_2 \mid x_3, \dots, x_n, C_k) \dots p(x_{n-1} \mid x_n, C_k) p(x_n \mid C_k) p(C_k) \end{aligned}$$

چون ویژگی‌های باینری هستند پس برای هر احتمال ۲ به توان تعداد متغیرها حالت در نظر بگیریم، بنابرین تعداد پارامترها برای حالت باینری:

$2+4+\ldots+2^{n+1} = 2^{n+2}-1-1=2(2^{n+1}-1)$

ب)

اگر ویژگیها با دانستن دسته، از یکدیگر مستقل باشند به مدل بیز ساده می‌رسیم:

$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,. \end{aligned}$$

که در آن تعداد پارامترهای به صورت زیر است:

2+4+…+4=2+4n=2(2n+1)

ج)

در این حالت نیز مانند قسمت ب عمل می‌کنیم با این تفاوت که پارامترها به گونه دیگری هستند مثلا اگر همه ویژگی‌ها نسبت به یک کلاس دارای m پارامتر باشند داریم:

$$2+2m+\ldots+2m=2+2mn=2(mn+1)$$

$$P(B|D=T) = P(B=T|D=T) + P(B=F|D\neq T)$$

$$= \frac{P(D=T|B=T)\,P(B\neq T)}{P(D=T)} + \frac{P(D\neq T|B=F)\,P(B=F)}{P(D=T)}$$

$$= \frac{.241 \times .242 + .8772 \times .758}{.769966} = .939313683980$$

$$P(D=T) = P(D=T|B=T, C=T)\,P(B=T, C=T) + P(D=T|B=T, C=F)\,P(B=T, C=F)$$
$$+ P(D=T|B=F, C=F)\,P(B=F, C=F) + P(D=T|B=F, C=T) \times P(B=F, C=T)$$

$$\underline{\underline{P(B, C)}} = P(B, C|A=T)\,P(A=T) + P(B, C|A=F)\,P(A=F)$$

$$B \perp C \,|\, A \Rightarrow \underline{\underline{P(B,C)}} = P(B|A=T)\,P(C|A=T)\,P(A=T)$$
$$+ \;\; '' \;\; '' \;\; F \;\; '' \;\; '' \;\; F \;\; '' \;\; '' \;\; F$$

$$\underline{\underline{P(C)}} = P(C|A=T)\,P(A=T) + P(C|A=F)\,P(A=F)$$

$$\underline{\underline{P(B)}} = P(B|A=T)\,P(A=T) + P(B|A=F)\,P(A=F)$$

$$\underline{\underline{P(D|B)}} = P(D|B, C=T)\,P(C=T) + P(D|B, C=F)\,P(C=F)$$

احتمال هایی که زیر آن ها خط کشیده شده است قواعد کلی هستند و برای محاسبه نهایی استفاده می شوند ⇐

$$P(B=F) = .63 \times .2 + .79 \times .8 = .758 \Rightarrow P(B=T) = .242$$

$$P(D=T|B=T) = .5 \times .26 + .15 \times .74 = .241$$

$$P(C=T) = .3 \times .2 + .25 \times .8 = .26 \Rightarrow P(C=F) = .74$$

$$P(D\neq T|B=F) = .67 \times .26 + .95 \times .74 = .8772$$

$$P(D=T) = .5 \times .0642 + .15 \times .1778 + .67 \times .0318 + .95 \times .7262 = .769966$$

$$P(B=T, C=T) = .3 \times .2 \times .37 + .25 \times .21 \times .8 = 0.0642$$

$$P(B=T, C=F) = .37 \times .7 \times .2 + .21 \times .75 \times .8 = .1778$$

$P(B=F, C=T) = 1 -$ بقیه $= .0318$

$$P(B=F, C=F) = .7 \times .63 \times .2 + .79 \times .75 \times .8 = .7262$$

(۴

منظور از مدل Generative ، هر نوع فرایند یادگیری طبقه‌بندی با استفاده از تخمین احتمال مشترک P(x,y) یا هر نوع فرایند یادگیری طبقه‌بندی است که با استفاده از تخمین احتمال مقدم، که در آن y یک طبقه و x توصیفی شئی است که باید طبقه‌بندی شود. با توجه به این مدل یا تخمین‌ها، می‌توان اشیای ترکیبی از توزیع مشترک ایجاد کرد. مدل Generative در مقابل مدل Discriminative است که در آن مدل یا تخمینی از P(y|x) بدون ارجاع به تخمینی صریح از P(x)، P(y,x) یا P(y|x) ایجاد می‌شود.

همچنین دسته‌بندی به عنوان رویکردهای افتراقی بر اساس تابع تصمیم نیز مرسوم است که به صورت مستقیم از ورودی x به خروجی y نگاشت می‌کند (مثل ماشین‌های بردار پشتیبان، شبکه‌های عصبی و درخت‌های تصمیم)، که در آن ریسک تصمیم بدون تخمین P(y,x)، P(y|x) یا P(y|x) کمینه می‌شود.

مثال استاندارد از مدل Generative، نایو بیز و از مدل Discriminative لاجستیک رگرسیون است.

مدل Generative هنگامی که نمونه‌ها کمیاب هستند خوب عمل می‌کند در حالی که مدل Discriminativ عملکرد خطای مجانبی بهتری دارد.

(۵

مدل Naive Bayes یک مدل Generative است در صورتی که مدل Logistic Regression یک مدل Discriminative است.

به این معنی که مدل Bayes پارامترهای P(Y) and P(X|Y) را تخمین می‌زند، در صورتی که Logit پارامترهای P(Y|X) را تخمین می‌زند.

بنابرین مدل Logit پیشفرضی بر روی توزیع داده ندارد و فقط سعی می‌کند MCLE احتمال P(Y|X) را بیشینه کند.
همچنین فرض استقلال شرطی در Logit به اندازه Bayes محکم و صریح نیست و Logit می‌تواند پارامترهایی متفاوت در این مواقع داشته باشد.

گر چه، اگر پارامترهای مدل Logit را با استفاده از مدل بیز ساده حساب شود، دو مدل یک دسته‌بند را ایجاد می‌کنند.

برای آموزش Logit به O(n) داده نیاز دارد(n تعداد ویژگی) در حالی که Bayes با O(lgn) داده قابل آموزش است و در عمل نشان داده شده است زمانی که داده زیاد است Logit بهتر عمل می‌کند و زمانی که داده کم باشد Bayes دسته بند بهتری ایجاد می‌کند (Generalization بهتری دارد).

مدل Bayes ، بایاس بیشتر و واریانس کمتری نسبت به مدل Logit دارد، ولی اگر تعداد داده‌ها به بی‌نهایت میل کند هر دو دسته‌بند برابر هم خواهند شد

پیاده‌سازی

Logistic Regression

الف)

کد در فایل main_lr.ipnb موجود است. در تمامی قسمت‌ها از ضریب یادگیری ۰.۰۱ استفاده شد.

ب)

لیست مقادیر خطای ۱۰-fold برای لامبداهای مختلف به صورت زیر به دست آمد:

```
Lambda          Error
 [ 2.           0.25597484]
 [ 1.           0.25974843]
 [ 0.           0.25597484]
 [-1.           0.25786164]
 [-2.           0.25660377]
 [-3.           0.26100629]
 [-4.           0.26100629]
 [-5.           0.26100629]
 Picking best lambda --> 2
```

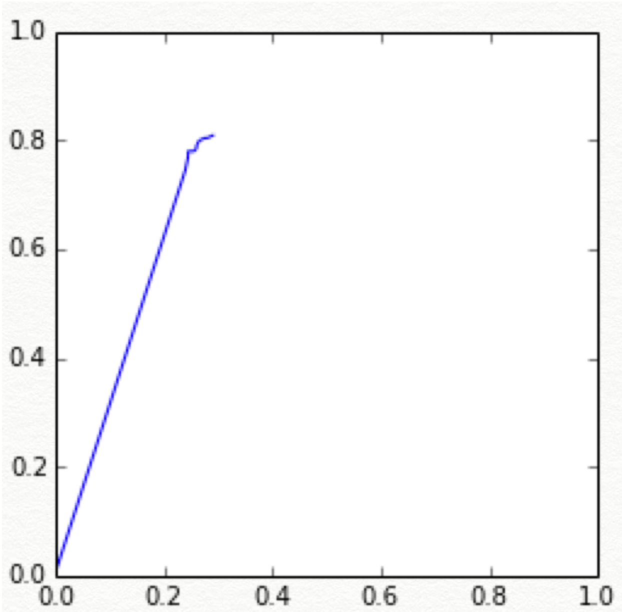
بنابرین لامبدا بهینه ۱۰ به توان ۲(است. تعداد قدم‌ها برای هر مرحله ۲۰۰۰ قدم بود.)

ج)

در این قسمت confusion matrix و خطای داده آموزشی و آزمایشی برای لامبدا ۱۰ به توان ۲ و با ۱۰۰۰۰ قدم به صورت زیر محاسبه گردید:

```
Test confusion matrix:
[[ 146.   45.]
 [  50.  159.]]
Test error:
0.2375
Train confusion matrix:
[[ 637.  149.]
 [ 167.  647.]]
Train error:
0.1975
```

نمودار ROC نیز با جابجا کردن T به صورت زیر به دست آمد:



Naïve Bayes

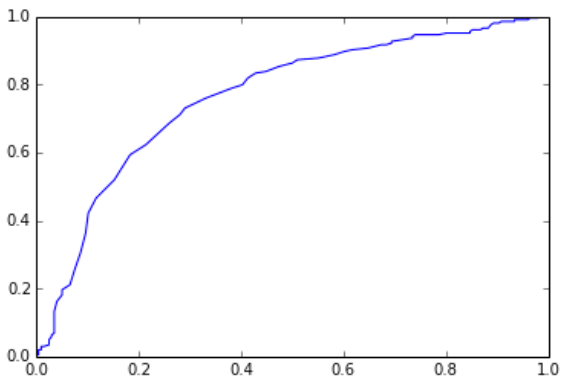
الف)

کد در فایل main_naive.ipnb موجود است.

Test error	0.240625
Train error	0.295

ب)

برای به دست آوردن نمودار ROC ابتدا تفاضل احتمال‌های پسین محاسبه شد و با قرار داده T از مینیمم تا ماکسمیم مقدار تفاضل، specifity و ۱-sensivity محاسبه گردید:



با توجه به نمودار ROC و AUC متناظر با آن این مدل نسبتا خوب است و خیلی بهتر از یک مدل تصادفی عمل کرده است.

Bayes Network

الف) کد این قسمت در فایل main_bayes_net.ipnb موجود است.

مدل‌های بیز ساده زیر استفاده شد:

Model\Accuracy	
Naive1	0.854221
Naive2	0.689259
Naive3	0.814276

هر مدل بیز ساده با حذف یکی از ویژگی‌ها ساخته شد.

ب)

ویژگی‌های موجود برای این مجموعه داده به صورت زیر مشخص شده است:

class values:

unacc, acc, good, vgood

attributes:

buying: vhigh, high, med, low.

maint: vhigh, high, med, low.

doors: 2, 3, 4, 5more.

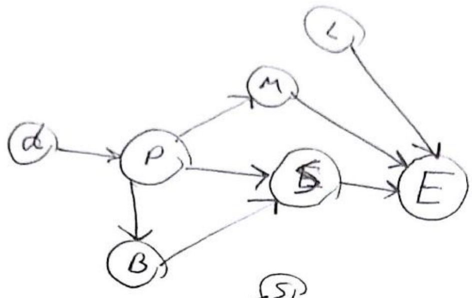
persons: 2, 4, more.

lug_boot: small, med, big.

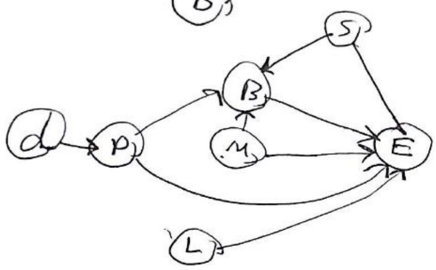
safety: low, med, high.

با توجه به ویژگی‌های بالا گراف مدل‌ها در تصویر زیر آورده شده است:

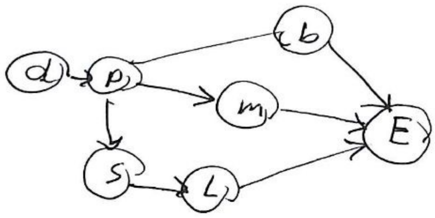
(هر گره شبکه بیزین اول حرف ویژگی متناظر است)



BaysNet1



BaysNet2



BaysNet3

توجیح مدل‌ها در زیر آورده شده است:

Model\Reason	**بعضی از فرض‌ها**
BayesNet1	فرض شده است که تعداد سرنشین ماشین به تعداد درهای ماشین وابستگی دارد. امنیت به تعداد سرنشین وابسته است. امنیت ماشین به قیمت ماشین وابستگی دارد.
BayesNet2	قیمت ماشین به امنیت ماشین وابستگی دارد. قیمت ماشین به تعداد سرنشین وابسته است. قیمت ماشین به هزینه نگهداری ماشین وابسته است.
BayesNet2	یک فرض این قسمت وابستگی اندازه جای پا به امنیت ماشین است، یعنی هر چه جای پا بیشتر باشد احتمالا ماشین ایمن‌تر بوده است!

دقت مدل‌ها در جدول زیر آمده است:

Model\Accuracy	
BayesNet1	0.671881
BayesNet2	0.939824
BayesNet3	0.697933

مشاهده می‌شود که ساختار دوم بالاترین دقت را داشته است. بنابرین وابستگی بین ویژگی‌ها بسیار خوب در نظر گرفته شده است. . بنابرین برای داده‌های ناشناخته استخراج ویژگی بهتری صورت گرفته است.