به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس یادگیری ماشین

استاد ناظرفرد

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: پرسش‌های تشریحی

سوال ۱

الف) اگر مقدار $k = 1$ باشـــد، وقتی که داده validation هر یک از چهار داده کلاس + باشـد، مدل به اشـتباه آن را – پیش‌بینی می‌کند چراکه داده – در میان آن‌ها فاصـله کمتری تا هر چهار داده دیگر نسبت به همدیگر دارد (فاصله مرکز مربع تا رئوس کمتر از فاصـــله دو راس مربع اســت. بدهی اســت که خود این داده منفی هم به اشـتباه دسته‌بندی خواهد شد. داده‌ها منفی طرف دیگر به وضوح درست دسته‌بندی می‌شوند.

باتوجه به اینکه k نباید مقدار زوج داشـته باشـد، حال باید$k = 3$ را بررسـی کنیم. در این حالت اتفاقی که برای چهار داده + رخ داد دیگر اتفاق نمی‌افتد چون به هر داده + دو داده + و یک داده – نزدیک اســـت اما همچنان داده – میان آن‌ها به اشـــتباه دسته‌بندی می‌شود. داده‌های منفی طرف دیگر به وضوح درست دسته‌بندی می‌شوند.

اگر $k = 5$ باشـد، برای هر یک از داده‌های + همچنان سـه داده + نزدیک‌تر وجود دارد که در نتیجه این‌ها به درسـتی ارزیابی می‌شـوند. برای داده‌های – در طرف چپ هم به وضوح نتیجه دسته‌بندی درست خواهد بود. اما همچنان داده – میان داده‌های مثبت به چهار داده + نزدیک است و دسته‌بندی اشتباهی خواهد داشت.

اگر $k \geq 7$ باشد در این صورت برای هر یک از داده‌های + تنها سه داده + باقی می‌ماند و درنتیجه همگی به اشـــتباه به کلاس – تعلق می‌گیرند. در نتیجه متوجه می‌شـــویم برای این مسئله حالت $k = 3$ و $k = 5$ با یک خطا بهترین حالت است.

ب) مانند تعیین مقدار هر ابرپارامتر دیگر می‌توان از یک مجموعه validation برای تعیین آن استفاده کرد. یعنی آنکه به ازای مقادیر k مختلف بررسی می‌کنیم که خطای مجموعه validation به چه شـــکلی بوده اســـت. جایی که کمترین خطا را برای این مجموعه داشتیم احتمالا k مناسب داشته است.

سوال ۲

الگوریتم KNN را می‌توان یک الگوریتم تمایزگر دانست؛ چراکه در این الگوریتم حساب می‌شود که یک داده به کدام کلاس تعلق دارد و مرزهای کلاس‌ها مختلف تشکیل می‌شود (مانند سلول‌های voronoi برای 1NN) و نمی‌توان الگوی یک کلاس را محاسبه کرد و توسط آن داده‌های جدید ایجاد کرد.

الگوریتم درخت تصمیم را هم باید یک الگوریتم تمایزگر دانست؛ چراکه در این الگوریتم هم باز صرفا بررسی می شود که یک داده به کدام کلاس تعلق دارد و می‌توان مرزها را برای کلاس‌های مختلف تشکیل داد ولی نمی‌توان توزیع هر کلاس را بدست آورد.[۲۱]

سوال ۳

الف)

$$\sigma(a) = \frac{1}{1+e^{-a}} \rightarrow \frac{d\sigma(a)}{da} = \frac{e^{-a}}{(1+e^{-a})^2} = \sigma(a) * \left(\frac{e^{-a}}{1+e^{-a}}\right) = \sigma(a) * \left(1 - \frac{1}{1+e^{-a}}\right)$$
$$= \sigma(a) * \big(1 - \sigma(a)\big)$$

ب)

$$\hat{y}^{(i)} = p\big(C_1|x^{(i)}\big) = \sigma\big(w^T x^{(i)}\big)$$

$$????f = -\log \sigma(w^T x)$$

ج)

$$??? \frac{\partial f}{\partial w} =$$

د)؟

---

[1] https://stats.stackexchange.com/questions/105979/is-knn-a-discriminative-learning-algorithm

[2] https://stats.stackexchange.com/questions/12421/generative-vs-discriminative

سوال ۴

الف)

$$p(buy = yes) = \frac{9}{14}\ , p(buy = no) = \frac{5}{14}$$

$$p(buy = yes \mid X_1) = \frac{p(X_1 \mid buy = yes) * p(buy = yes)}{p(X_1)}$$

$$\begin{aligned} p(X_1 | buy = yes) \\ &= p(age = youth \mid buy = yes) * \ p(income = high | \ buy = yes) \\ &* \ p(student = yes | \ buy = yes) * \ p(credit = fair \mid buy = yes) \\ &= \frac{2}{9} * \frac{2}{9} * \frac{6}{9} * \frac{6}{9} \end{aligned}$$

$$p(buy = no \mid X_1) = \frac{p(X_1 \mid buy = no) * p(buy = no)}{p(X_1)}$$

$$\begin{aligned} p(X_1 | buy = no) \\ &= p(age = youth \mid buy = no) * \ p(income = high | \ buy = no) \\ &* \ p(student = yes | \ buy = no) * \ p(credit = fair \mid buy = no) \\ &= \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * \frac{2}{5} \end{aligned}$$

$$p(buy = yes \mid X_1) \ \approx \frac{0.014}{p(X_1)}\ , p(buy = no | X_1) \approx \frac{0.006}{p(X_1)}$$

$$\rightarrow \ p(buy = yes \mid X_1) > \ p(buy = no \mid X_1)$$

$$p(buy = yes \mid X_2) = \frac{p(X_2 \mid buy = yes) * p(buy = yes)}{p(X_2)}$$

$$\begin{aligned} (X_2 | buy = yes) \\ &= p(age = senior \mid buy = yes) * \ p(income = low | \ buy = yes) \\ &* \ p(student = no | \ buy = yes) * \ p(credit = excellent \mid buy = yes) \\ &= \frac{3}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} \end{aligned}$$

$$p(buy = no \mid X_2) = \frac{p(X_2 \mid buy = no) * p(buy = no)}{p(X_2)}$$

$$(X_2 | buy = no)$$
$$= p(age = senior \mid buy = no) * p(income = low \mid buy = no)$$
$$* p(student = no \mid buy = no) * p(credit = excellent \mid buy = no)$$
$$= \frac{2}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5}$$

$$p(buy = yes \mid X_2) \approx \frac{0.007}{p(X_2)}, p(buy = no | X_2) \approx \frac{0.013}{p(X_2)}$$

$$\rightarrow p(buy = yes \mid X_2) < p(buy = no \mid X_2)$$

$$p(buy = yes \mid X_3) = \frac{p(X_3 \mid buy = yes) * p(buy = yes)}{p(X_3)}$$

$$(X_3 | buy = yes)$$
$$= p(age = middle \mid buy = yes) * p(income = medium \mid buy = yes)$$
$$* p(student = no \mid buy = yes) * p(credit = fair \mid buy = yes)$$
$$= \frac{4}{9} * \frac{4}{9} * \frac{3}{9} * \frac{6}{9}$$

$$p(buy = no \mid X_3) = \frac{p(X_3 \mid buy = no) * p(buy = no)}{p(X_3)}$$

$$(X_3 | buy = no)$$
$$= p(age = middle \mid buy = no) * p(income = medium \mid buy = no)$$
$$* p(student = no \mid buy = no) * p(credit = fair \mid buy = no) = 0$$

$$p(buy = yes \mid X_3) > p(buy = no | X_3)$$

ب)

ابتدا باید بررسی کرد که برای ریشه درخت کدام ویژگی مناسب‌تر است.

$$E(age = youth) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} \approx 0.442 + 0.528 = 0.97$$

$$E(age = middle) = -1\log 1 = 0$$

$$E(age = senior) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} \approx 0.442 + 0.528 = 0.97$$

$$Gain(age) = E(S) - \left(\frac{5}{14} * 0.97 + \frac{5}{14} 0.97 + \frac{4}{14} * 0\right) = E(s) - 0.692$$

$$E(income = high) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = 1$$

$$E(income = medium) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} \approx 0.389 + 0.528 = 0.917$$

$$E(income = low) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.311 + 0.5 = 0.811$$

$$Gain(income) = E(S) - \left(\frac{4}{14} * 1 + \frac{6}{14} * 0.917 + \frac{4}{14} * 0.811\right) = E(s) - 0.910$$

$$E(student = yes) = -\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7} \approx 0.190 + 0.401 = 0.591$$

$$E(student = no) = -\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7} \approx 0.523 + 0.461 = 0.984$$

$$Gain(Student) = E(S) - \left(\frac{7}{14} * 0.591 + \frac{7}{14} * 0.984\right) = E(S) - \ 0.787$$

$$E(credit = fair) = -\frac{2}{8}\log\frac{2}{8} - \frac{6}{8}\log\frac{6}{8} \approx 0.5 + 0.311 = 0.811$$

$$E(credit = excellent) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$$

$$Gain(credit) = E(S) - \left(\frac{8}{14} * 0.811 + \frac{6}{14} * 1\right) = E(s) - 0.892$$

به نظر می‌رسد برای ریشه پرسش راجع به age بهترین گزینه است.

حال باید بررسی کرد که برای شاخه age = youth کدام ویژگی مناسب‌تر است.

age	income	student	credit	Buy
Youth	High	No	Fair	**NO**
Youth	High	No	Excellent	**NO**
Youth	Medium	No	Fair	**NO**
Youth	Low	Yes	Fair	**YES**
Youth	Medium	Yes	Excellent	**YES**

طبیعتا برای این شـاخه هم می‌توان مشـابه با محاسـبات گره ریشـه پیش رفت؛ اما پیش از آن توجه کنید که اگر student مورد سـوال قرار بگیرد قطعا دو گره خالص ایجاد خواهد شـد این درحالی اسـت که در مابقی ویژگی‌ها حداقل یک گره ناخالص باقی می‌ماند. (برای income، مقدار medium؛ برای credit، مقدار fair) پس قطعا Gain ویژگی Student برابر بیشـینه مقدار که انتروپی همین گره age= youth باشـد است که از مابقی بیشتر است و نیازی به محاسبات بیشتر نیست.

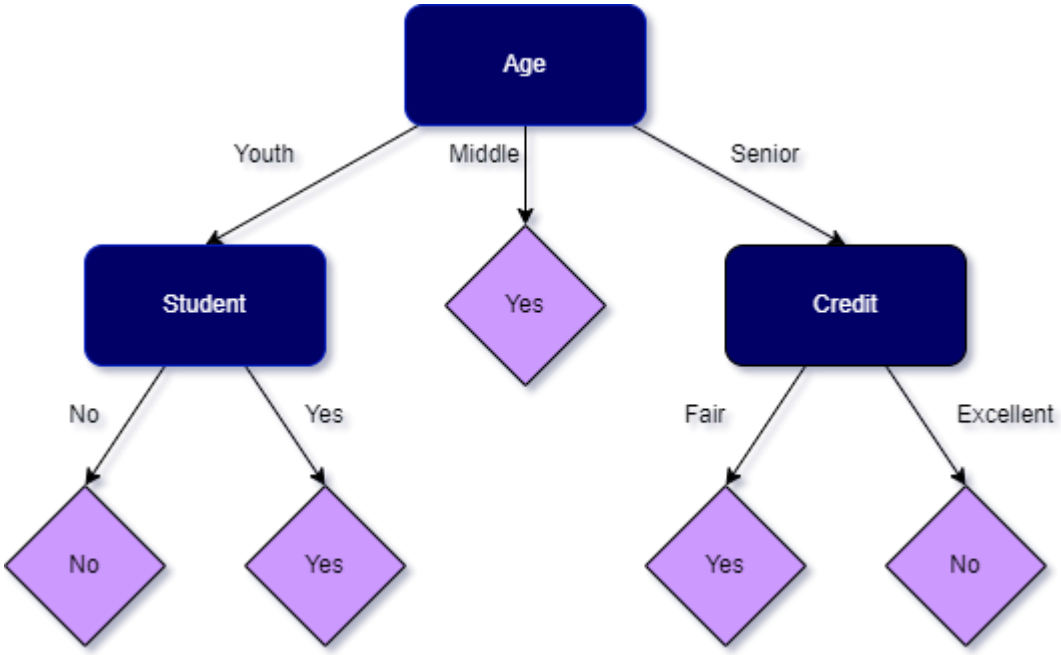
برای شـاخه age = middle یک گره خالص باقی می‌ماند که نیاز به ادامه ندارد و می‌دانیم به پاسخ yes خواهیم رسید.

age	income	student	credit	Buy
Middle	High	No	Fair	**YES**
Middle	Low	Yes	Excellent	**YES**
Middle	Medium	No	Excellent	**YES**
Middle	High	Yes	Fair	**YES**

برای شـاخه آخر یعنی age = senior، واضـحا و مشـابه با بحث گره age =youth انتخاب ویژگی credit بهترین گزینه است چراکه دو گره خالص حاصل می‌شود و Gain این ویژگی برابر با آنتروپی گره age = senior اسـت که بیشـترین Gain ممکن اسـت. برای دو ویژگی دیگر هم مشخص است که گره‌های حاصل تماما خالص نخواهند بود.

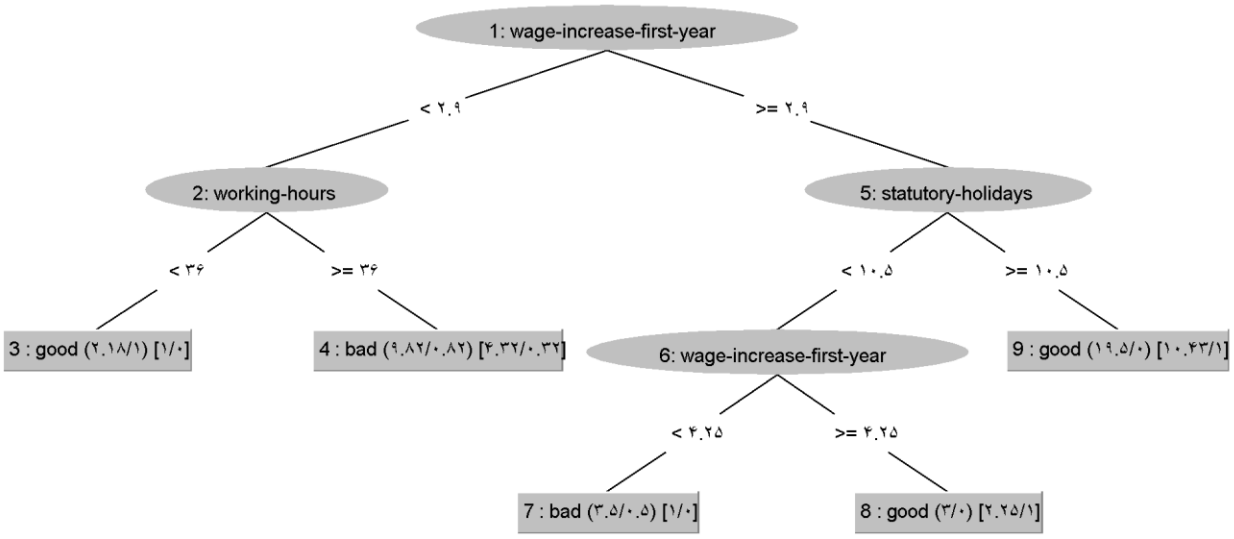
age	income	student	credit	Buy
Senior	Medium	No	Fair	**YES**
Senior	Low	Yes	Fair	**YES**
Senior	Low	Yes	Excellent	**NO**
Senior	Medium	Yes	Fair	**YES**
Senior	Medium	No	Excellent	**NO**

درخت نهایی برابر خواهد بود با:



سوال ۵

الف) برای این سوال از الگوریتم REPTree موجود در نرم‌افزار استفاده کردم و سایر پارامترها را تغییر ندادم.

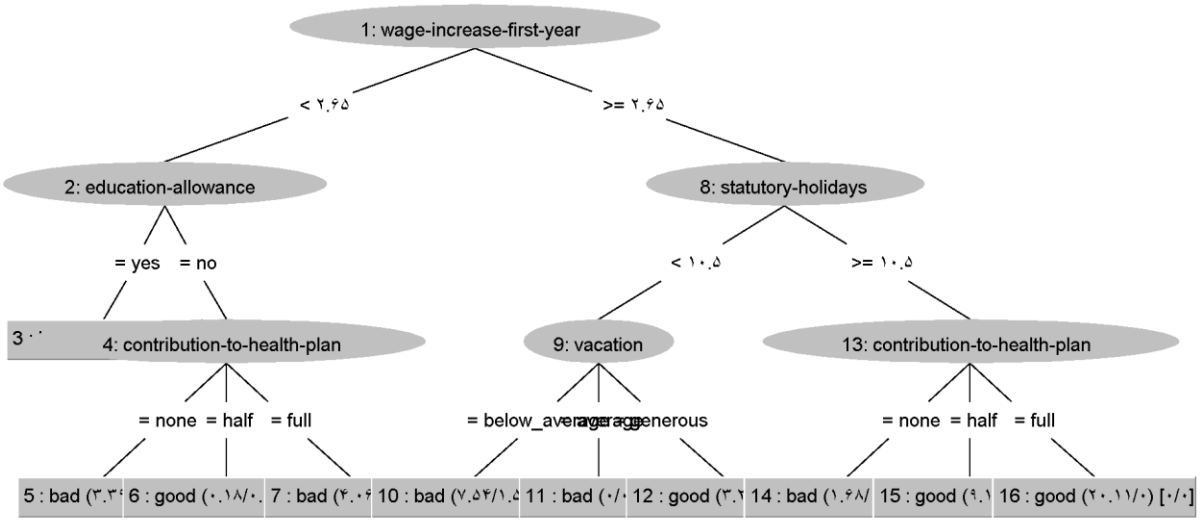


ب)

```
=== Confusion Matrix ===

  a  b   <-- classified as
 ۱۴  ۶ |  a = bad
  ۶ ۳۱ |  b = good
```

ج)



د) مطابق انتظار درخت هرس‌نشـــده تعداد گره بیشـــتری دارد. همچنین می‌توان مشاهده کرد که درصد خلوص در گره‌های انتهایی درخت هرس‌نشده بیشتر است.

بخش دوم: پرسش‌های پیاده‌سازی

سوال ۱

الف)

ب)

ج)

سوال ۲

برای پیاده‌سـازی این سـوال، از پیاده‌سـازی سـابق خودم در دوره کارشـناسـی اسـتفاده مجدد کردم.[۳]

در عین حال توجه کنید که مطابق اعلام تدریس‌یاران برای پیاده‌ســـازی این ســـوال از کتابخانه آمده برای درخت تصمیم استفاده کردم.

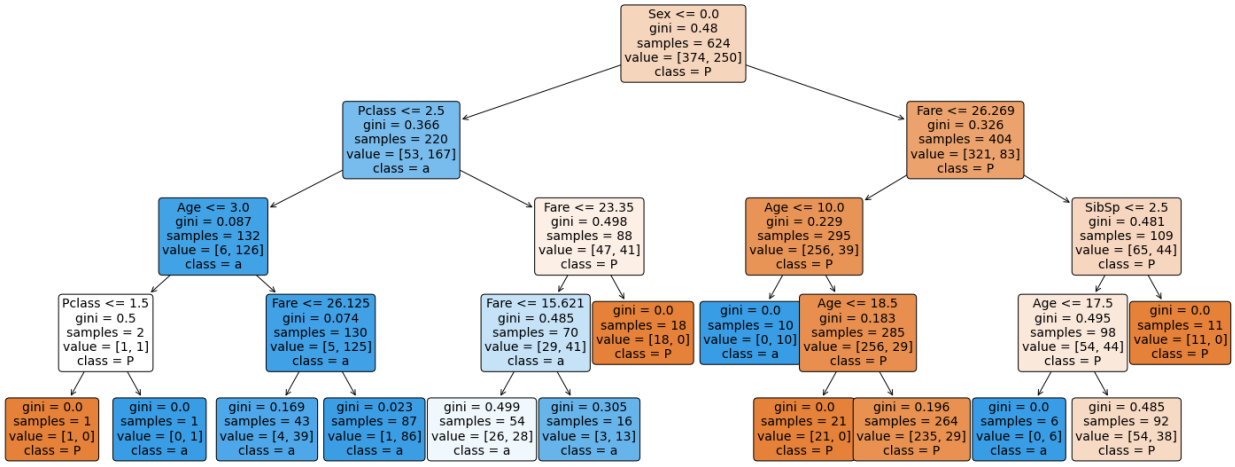
نهایتا توجه کنید که برای تعیین پارامترهای مدل مجموعه آموزشـــی را به دو مجموعه آموزشی و validation شکستم. (۷۰٪ برای train و ۳۰٪ برای validation)

الف) ابتدا قبل از تصمیم برای مدیریت مقادیر گم‌شده، باید ببینیم کدام ویژگی‌ها باید حذف شود. در این مجموعه‌داده ویژگی‌های PassengerID، Name و Ticket به نوعی نقش ID دارند و باید حذف شوند. همچنین ویژگی Cabin مقادیر گم‌شده زیادی دارد و در عین حال چندان مفید نیست. برای سایر ویژگی‌ها عددی مانند Age از میانگین

[3] https://gitlab.com/aut-data-mining/titanic/

همان ویژگی برای سایر داده‌ها استفاده کردیم. برای ویژگی Embarked هم تنها دو داده بدون مقدار بودند که برای آن‌ها بدون دلیل یک مقدار تصادفی درنظر گرفتیم.

ب) درخت تصمیم (نسبتا) بهینه را می‌توانید در تصویر زیر مشاهده کنید. عمق درخت ۴ و معیار خلوص Gini است. پس از آموزش داده‌های آموزش و Validation را پیش‌بینی کردیم و در فایل train_predicts.csv ذخیره کردیم.

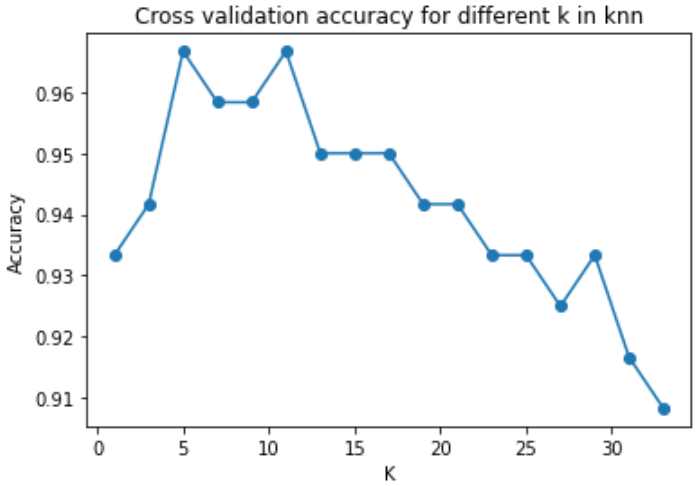


ج) مجموعه تست پردازش‌شده که شامل مقادیر پیش‌بینی شده هستند در فایل test_predicts.csv آورده شده است.

سوال ۳

ابتدا باید اشاره کنم برای این مسئله ابتدا داده‌ها را shuffle کردم و سپس ۸۰٪ داده‌ها را برای آموزشی و ۲۰٪ آن را برای تست در نظر گرفتم. توجه کنید که الگوریتم Cross Validation روی قسمت‌های مختلف مجموعه آموزشی اعمال شده است.

الف) در نمودار زیر میزان صحت به ازای k های مختلف آورده شده است. به نظر می‌رسد مقدار k برابر با ۵ یا ۱۱ بهترین نتیجه را داشته است. برای قسمت بعد مقدار k را برابر با ۱۱ درنظر می‌گیرم. در اینجا لازم است این نکته را متذکر شوم که با بررسی‌ها انجام شده و shuffle کردن‌های متنوع امکان تغییر k پیشنهادی وجود دارد!



ب) برای مجموعه آموزشی میزان صحت برابر با ۹۶/۶۷٪ است و برای مجموعه تست میزان صحت برابر با ۹۳/۳۳٪ است. در دو ماتریس درهم‌ریختگی زیر می‌توانید ارزیابی دقیق‌تری داشته باشید.

ماتریس درهم‌ریختگی مجموعه آموزشی

	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	۳۹	۰	۰
Iris-versicolor	۰	۴۱	۱
Iris-virginica	۰	۳	۳۶

ماتریس درهم‌ریختگی مجموعه تست

	Iris-setosa	Iris-versicolor	Iris-virginica
Iris-setosa	۱۱	۰	۰
Iris-versicolor	۰	۷	۱
Iris-virginica	۰	۱	۱۰