پاییز ۱۳۹۸

یادگیری ماشین

تمرین دوم

پرهام الوانی

۹۸۱۳۱۹۱۰

بخش اول

۱. در صورتی که درخت هرس شود از ارتفاع آن کاسته می‌شود و بنابراین از بیش برازش جلوگیری می‌شود. در زمان ساخت درخت تصمیم گاهی ممکن است شاخه‌هایی تنها بر اساس یک داده ساخته شوند، در این صورت می‌توان با حرص کردن این شاخه‌ها را حذف کرده و از بیش برازش جلوگیری کرد.

۲.

الف) اینطور به نظر می‌رسد که ویژگی B برای دسته‌بندی مناسبتر است.

ب)

$$-\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right)=1$$

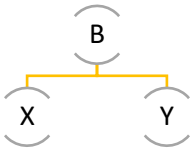
$$Gain(S,A)=1-\left(\frac{3}{4}*\left(-\frac{2}{3}\log\left(\frac{2}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)\right)+\frac{1}{4}*(0)\right)$$

$$Gain(S,B)=1-\left(\frac{2}{4}*(0)+\frac{2}{4}*(0)\right)=1$$

$$Gain(S,C)=1-\left(\frac{2}{4}*\left(-\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{2}\log\left(\frac{1}{2}\right)\right)+\frac{2}{4}*\left(-\frac{1}{2}\log\left(\frac{1}{2}\right)-\frac{1}{2}\log\left(\frac{1}{2}\right)\right)\right)=0$$

بنابراین بهترین ویزگی همان ویژگی B می‌باشد که *Information Gain* بیشتری دارد.

پ)



۳. در روش جنگل تصادفی در زمان ساخت درخت، به جای انتخاب ویژگی که بیشترین تمایز را اعمال می‌کند، این ویژگی از بین یک زیرمجموعه تصادفی از ویژگی‌های انتخاب می‌شود. این الگوریتم با اینکه در نگاه اول خوب به نظر نمیرسد با این روش جلو *overfitting* را می‌گیرد.

Classification and Regression by

randomForest

Andy Liaw and Matthew Wiener

بخش دوم

۱. بهترین مقدار $k = 4$ می‌باشد. که در آن دقت الگوریتم برابر است با:

$$accuracy = \frac{5+5}{14} = \frac{5}{7}$$

۲. در الگوریتم k نزدیک‌ترین همسایه در صورتی که مقدار k افزایش پیدا کند اگوریتم بیشتر بایاس می‌شود و زمانی که مقدار k کم باشد اگوریتم واریانس بیشتری خواهد داشت. به این معنی در زمانی که مقدار k افزایش پیدا می‌کند از حالت overfitting به underfitting می‌رویم.

۳. در الگوریتم KNN با انتخاب مقدار بزرگ برای k می‌توان تاثیر داده‌های نویز را کاهش داد.

۴. در اشکال، مقدار k از سمت چپ به راست افزایش پیدا می‌کند چرا که مرز ناحیه‌ها در سمت چپ‌ترین شکل بیشتر از داده‌ها پیروی کرده است که نشان می‌دهد مقدار k کم می‌باشد.

بخش سوم

۱.

ج) بعد از بارگذاری مجموعه داده خواسته شده و رسم آن بر اساس ویژگی‌های خواسته شده داریم:

د) به ترتیب برای مقدار k برابر ۱ تا ۵ ماتریس‌های پریشانی زیر را داریم:

=== Confusion Matrix ===

a b c <-- classified as

50 0 0 | a = Iris-setosa

0 50 0 | b = Iris-versicolor

0 0 50 | c = Iris-virginica

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 50  0 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  3 47 |  c = Iris-virginica
```

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

=== Confusion Matrix ===

```
  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

هـ) برای اضافه کردن نویز از منوی پیش پردازش و فیلتر AddNoise استفاده می‌کنیم. با افزایش مقدار k از overfitting جلوگیری می‌شود بنابراین پیش‌بینی‌ها ممکن است روی داده‌های آموزش درست نباشند.

```
=== Confusion Matrix ===
 a  b  c   <-- classified as
52  0  0 |  a = Iris-setosa
 1 50  0 |  b = Iris-versicolor
 0  0 47 |  c = Iris-virginica
```

```
=== Confusion Matrix ===
 a  b  c   <-- classified as
51  0  1 |  a = Iris-setosa
 7 43  1 |  b = Iris-versicolor
 5  9 33 |  c = Iris-virginica
```

```
=== Confusion Matrix ===
 a  b  c   <-- classified as
43  4  5 |  a = Iris-setosa
 5 43  3 |  b = Iris-versicolor
 5  3 39 |  c = Iris-virginica
```

```
=== Confusion Matrix ===
 a  b  c   <-- classified as
43  4  5 |  a = Iris-setosa
 6 41  4 |  b = Iris-versicolor
 4  5 38 |  c = Iris-virginica
```

=== Confusion Matrix ===

 a b c <-- classified as

43 4 5 | a = Iris-setosa

 5 43 3 | b = Iris-versicolor

 4 5 38 | c = Iris-virginica

۲.

الف)

=== Confusion Matrix ===

  a   b <-- classified as

261   6 | a = democrat

  6 162 | b = republican

ب)

=== Confusion Matrix ===

  a   b <-- classified as

263   4 | a = democrat

  5 163 | b = republican

ج)

J48 pruned tree

----

physician-fee-freeze = n: democrat (253.41/3.75)

physician-fee-freeze = y

| synfuels-corporation-cutback = n: republican (145.71/4.0)

| synfuels-corporation-cutback = y

| | mx-missile = n

| | | adoption-of-the-budget-resolution = n: republican (22.61/3.32)

| | | adoption-of-the-budget-resolution = y

| | | | anti-satellite-test-ban = n: democrat (5.04/0.02)

| | | | anti-satellite-test-ban = y: republican (2.21)

| | mx-missile = y: democrat (6.03/1.03)

J48 unpruned tree

---

physician-fee-freeze = n

| adoption-of-the-budget-resolution = n

| | synfuels-corporation-cutback = n

| | | superfund-right-to-sue = n

| | | | el-salvador-aid = n

| | | | | religious-groups-in-schools = n: republican (2.01/1.0)

| | | | | religious-groups-in-schools = y: democrat (2.12/0.01)

| | | | el-salvador-aid = y: republican (2.01/1.0)

| | | superfund-right-to-sue = y: democrat (4.21/0.08)

| | synfuels-corporation-cutback = y: democrat (15.3/0.07)

| adoption-of-the-budget-resolution = y: democrat (227.75/1.57)

physician-fee-freeze = y

| synfuels-corporation-cutback = n

| | education-spending = n

| | | religious-groups-in-schools = n: republican (6.15/0.01)

| | | religious-groups-in-schools = y

| | | | duty-free-exports = n: republican (9.27/0.58)

| | | | duty-free-exports = y

| | | | | anti-satellite-test-ban = n: democrat (2.47/0.36)

| | | | | anti-satellite-test-ban = y: republican (2.03/0.0)

| | education-spending = y: republican (125.78/1.29)

| synfuels-corporation-cutback = y

| | mx-missile = n

| | | adoption-of-the-budget-resolution = n

| | | | immigration = n

| | | | | anti-satellite-test-ban = n

| | | | | | export-administration-act-south-africa = n

| | | | | | | handicapped-infants = n: democrat (3.97/1.97)

| | | | | | | handicapped-infants = y: republican (2.55/0.55)

| | | | | | export-administration-act-south-africa = y: republican (5.41/0.77)

| | | | | anti-satellite-test-ban = y: republican (2.04)

| | | | immigration = y: republican (8.63)

| | | adoption-of-the-budget-resolution = y

| | | | anti-satellite-test-ban = n: democrat (5.04/0.02)

| | | | anti-satellite-test-ban = y: republican (2.21)

| | mx-missile = y: democrat (6.03/1.03)

د) در صورتی که از هرس کردن استفاده کنیم ارتفاع درخت کاهش پیدا می‌کند.