K-Nearest Neighbor

예제 풀이

학습 목표

1. K-Nearest Neighbor 방법을 실제 데이터에 적용
2. 최근접 이웃의 개수 (parameter) 설정에 따라서 결과값이 달라질 수 있는가?
3. (hidden) 데이터를 나눌 때 어떠한 기준을 가지고 나눌 것인가?

축하합니다! 여러분은 고생 끝에 수원대학교 데이터과학부의 조교수로 임용 받았습니다!

데이터과학부는 한 학년에 110명이나 된다고 합니다!

110명은 서로 다른 성격과 특성을 가지고 있어서, 이 학생들을 어떻게 분류해서 수업을 진행할지 막막합니다.

이제 여러분은 3학년과 4학년, 총 220명을 맡아서 학생들의 학습 성취도를 학생의 성향에 맞는 방법으로 찾아서 분류하려고 합니다.

학습 성취도는 보통 시험과 프로젝트로 나뉜다고 합니다. 어떤 학생은 시험을 잘 보고 어떤 학생은 프로젝트를 잘 하는 거죠. 물론 둘 사이에는 어느정도 연관관계가 있어서 시험은 0점인데 프로젝트는 100점을 받는 학생은 없다고 가정합시다. 여러분은 최대한 학생들에게 유리하게 수업을 나누어 진행하려고 합니다.

그러고보니 수업을 진행하면서 알게 된 KNeighborsClassifier를 사용해보고 싶습니다.

주어진 데이터 세트를 사용하여 아래 그래프를 그리세요.

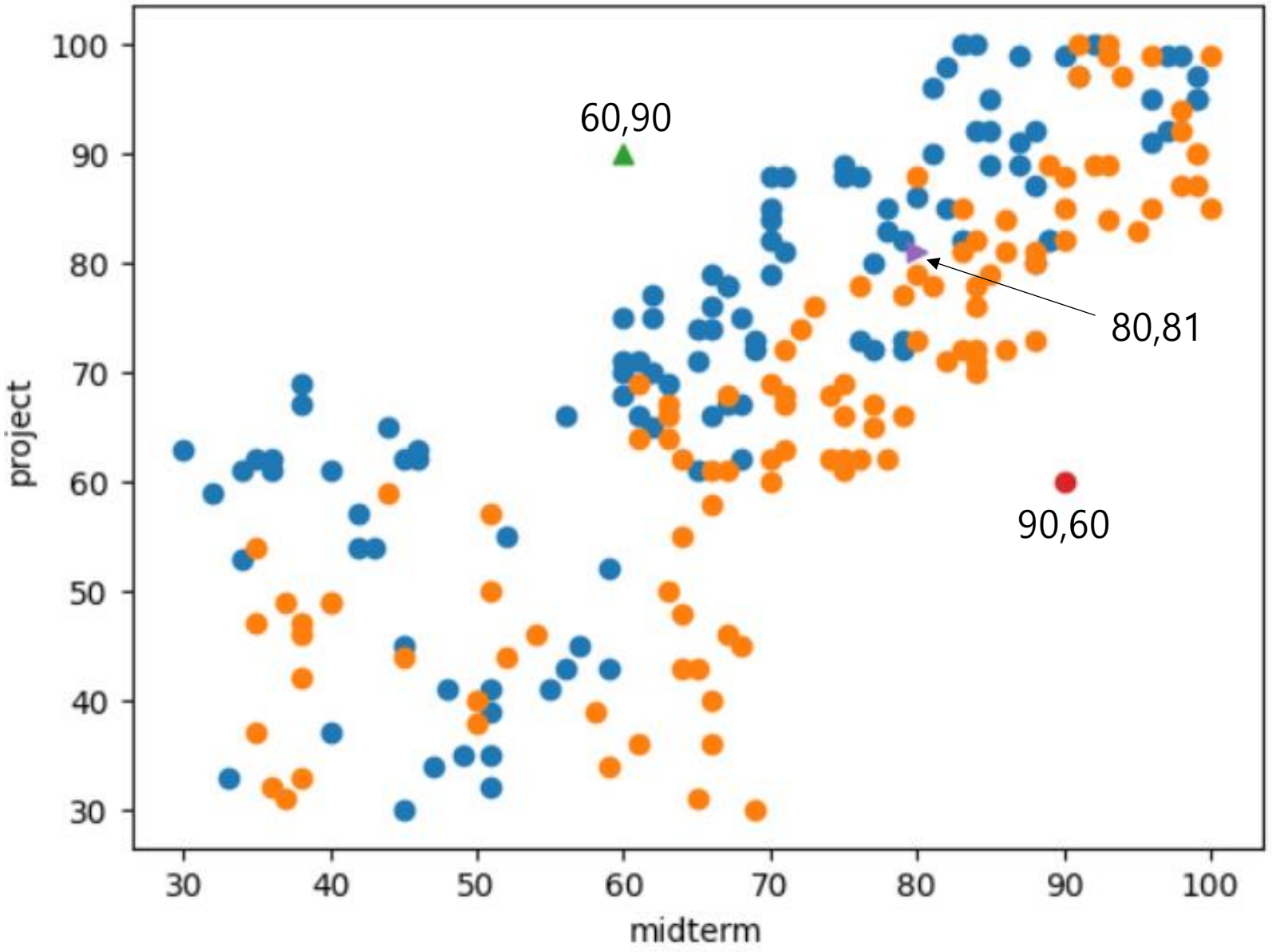
위 데이터 세트에서 최근접 이웃 데이터 3개를 사용하여서 모델을 훈련하고 예측 스코어를 구해보세요.
훈련한 모델을 사용하여 아래 학생의 학습 향상을 위해서는 어떤 학습 성취 방법을 사용해야 하는지 예측해보세요.

- 학생 1: 중간고사 60점, 프로젝트 90점
- 학생 2: 중간고사 90점, 프로젝트 60점
- 학생 3: 중간고사 80점, 프로젝트 81점

위 데이터 세트에서 최근접 이웃 데이터 30개를 사용하여서 모델을 훈련하고 예측 스코어를 구해보세요.
훈련한 모델을 사용하여 아래 학생의 학습 향상을 위해서는 어떤 학습 성취 방법을 사용해야 하는지 예측해보세요.

- 학생 1: 중간고사 60점, 프로젝트 90점
- 학생 2: 중간고사 90점, 프로젝트 60점
- 학생 3: 중간고사 80점, 프로젝트 81점

학습문제: https://github.com/mario2437/ML_lecture_1/blob/main/You%20are%20the%20professor



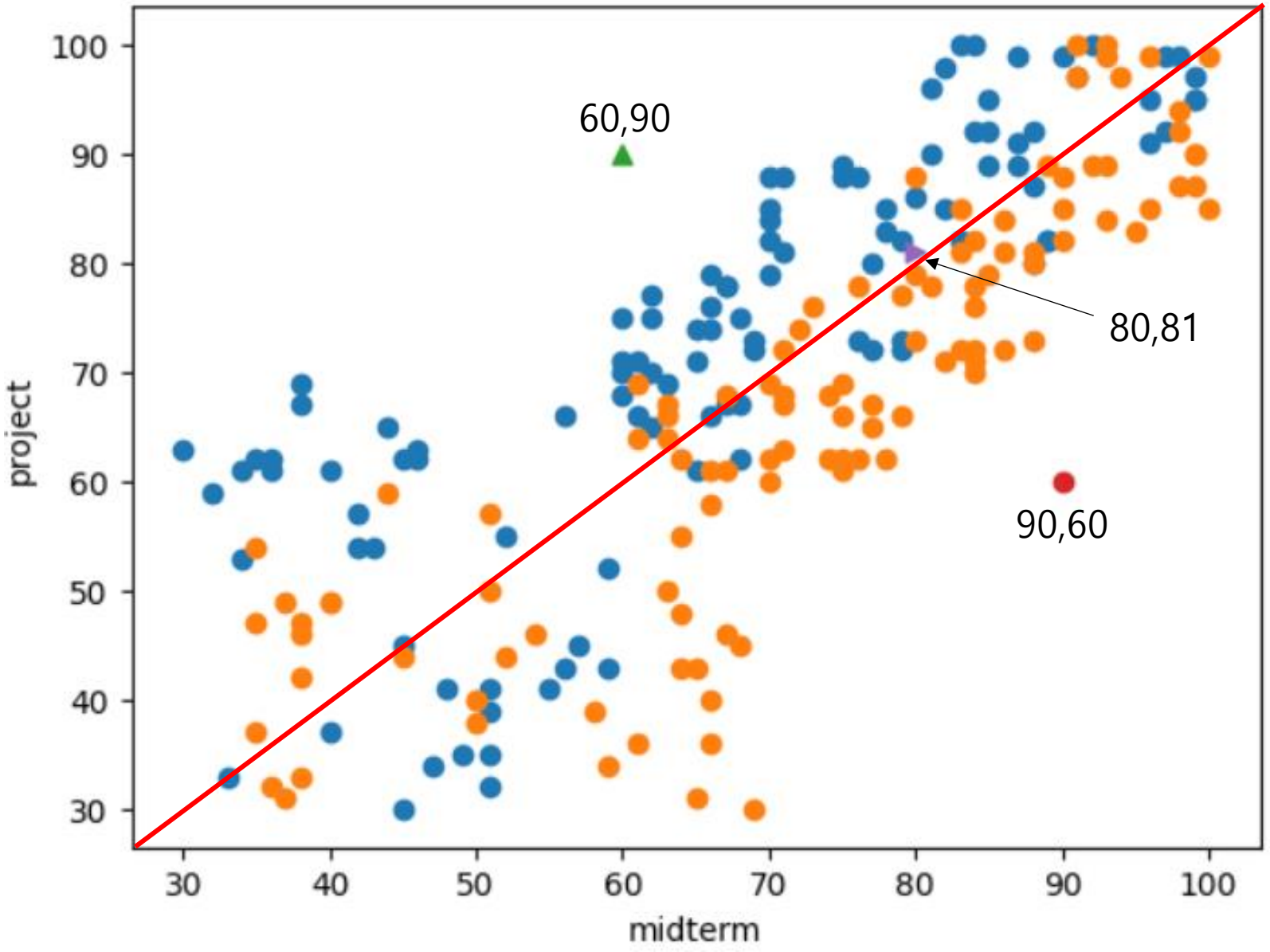
```
import matplotlib.pyplot as plt

plt.scatter(A_midterm, A_project)
plt.scatter(B_midterm, B_project)
plt.scatter(60, 90, marker = '^')
plt.scatter(90, 60, marker = 'o')
plt.scatter(80, 81, marker = '>')
plt.xlabel('midterm')
plt.ylabel('project')
plt.show()

# data trimming
midterm = A_midterm+B_midterm
project = A_project+B_project
test_data = [[l, w] for l, w in zip(midterm, project)]
```

```
# THIS IS TRICKY PART. CAN WE ASSIGN TEAM A AS "1" AND TEAM B AS "0". IF SO, WHAT IS YOUR REASON?
test_target = [1]*110 + [0]*110

# import KNeighborsClassifier model from the sklearn.neighbors
from sklearn.neighbors import KNeighborsClassifier

# apply KNeighborsClassifier using 3 nearest neighbors
kn3 = KNeighborsClassifier(n_neighbors=3)

# train the model
kn3.fit(test_data, test_target)
```

```
# calculate the prediction score
score3 = kn3.score(test_data, test_target)
print(score3)

a = kn3.predict([[60, 90]])
b = kn3.predict([[90, 60]])
c = kn3.predict([[80, 81]])

print(a)
print(b)
print(c)
```

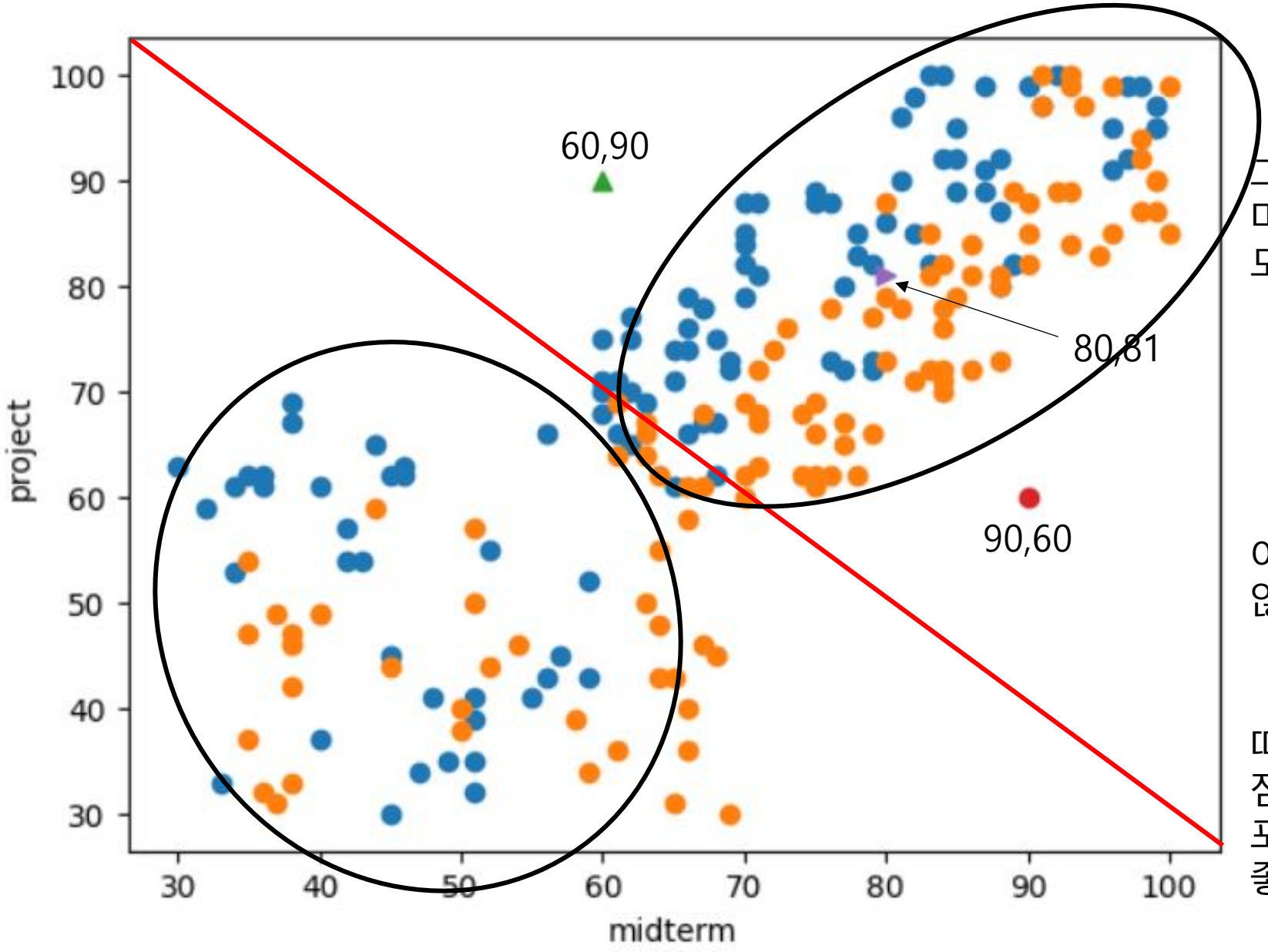
```
# apply KNeighborsClassifier using 30 nearest neighbors
kn30 = KNeighborsClassifier(n_neighbors=30)

# train the model
kn30.fit(test_data, test_target)

# calculate the prediction score
score30 = kn30.score(test_data, test_target)
print(score30)

d = kn30.predict([[60, 90]])
e = kn30.predict([[90, 60]])
f = kn30.predict([[80, 81]])

print(d)
print(e)
print(f)
```

Team B는 y=x 선을 기준으로 project 으로 편향되어 있음 -> 따라서 project로 annotation



Team A는 y=x 선을 기준으로 midterm 으로 편향되어 있음 -> 따라서 midterm 으로 annotation

BIG QUESTION
위 기준으로 분석하는 것이 합리적인가?



그러나 내가 제시한 기준은 위쪽 동그라미에 포함된 샘플 (> 30)에서는 어느정도 샘플링이 되지만,

아래쪽 동그라미에서는 샘플링이 되지 않고, 섞여 있음.

따라서, 중간고사를 80점, 프로젝트를 81점 받은 학생을 아래 동그라미 샘플이 포함되도록 parameter를 설정했다면, 좋지 못한 분석임.

학습 목표

1. K-Nearest Neighbor 방법을 실제 데이터에 적용
2. 최근접 이웃의 개수 (parameter) 설정에 따라서 결과값이 달라질 수 있는가?
3. (hidden) 데이터를 나눌 때 어떠한 기준을 가지고 나눌 것인가?

"지도학습"을 위해 데이터를 분류 하는 것은 특정한 "기준"이 필요함.

분석에 포함할 데이터와 기준을 정하는 것이 우리 데이터과학자들의 할 일.

따라서 이러한 기준으로 분석하는 것이 합리적인가?를 계속해서 질문해야 함.