

Intermediate Data Science

KNN, Linear Models and Regularization

Joanna Bieri DATA201

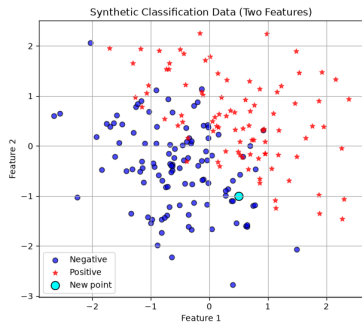
Mini Review - Supervised vs. Unsupervised

REVIEW

- In supervised learning you have a **labeled** data set, meaning that you have some examples of what the answers are.
- In unsupervised learning you have an **unlabeled** data set, meaning that you only have features or variables. You are trying to learn some underlying structure or patterns within the data.
- KNN, Linear and Logistic Regression are all supervised.

Mini Review - K-Nearest Neighbors

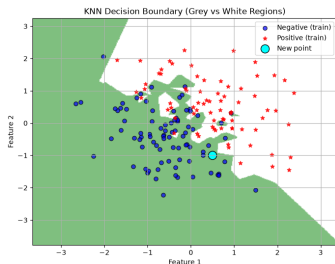
REVIEW



1. Choose a value for **K**.
2. Calculate the **distance** between the new data point and all training points.
3. Select the **K nearest neighbors**.
4. Classification: a **majority vote**. Regression: the **average**.

Mini Review - Is this a good fit?

REVIEW



```
knn = KNeighborsClassifier(n_neighbors=1)
knn.fit(X_train, y_train)
Accuracy on test set: 0.825
```

Mini Review - Plotting Training and Testing Accuracy

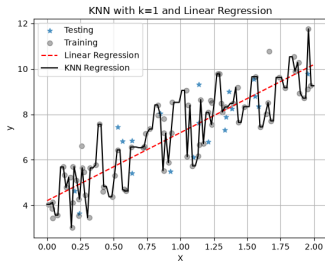
REVIEW



- **Underfitting:** the model does not do well on either the testing or training set.
- **Overfitting:** the model does really well on the training set but not very well on the testing set.

Mini Review - KNN Regression

REVIEW



Instead of having the neighbors vote, you take an average of the neighbors' values. With $k = 1$: test MSE 1.306, Linear Regression 0.918.

Mini Review - Scale the Data

REVIEW

StandardScaler rescales features to have mean = 0 and standard deviation = 1:

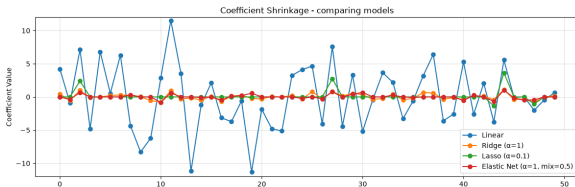
$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

```
scaler = StandardScaler()  
# find mu and sigma using the training data  
X_train_scaled = scaler.fit_transform(X_train)  
X_val_scaled = scaler.transform(X_val)
```

ALWAYS SCALE YOUR FEATURE DATA!!

Mini Review - Regularization

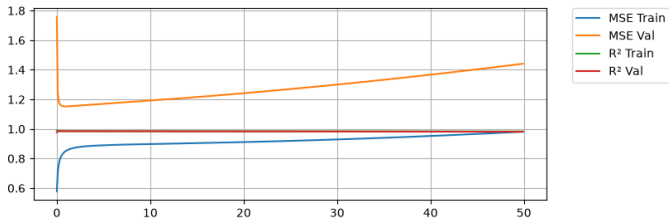
REVIEW



- **Ridge** (L2) adds $\alpha \sum b_j^2$. It keeps all features, the coefficients shrink toward zero.
- **Lasso** (L1) adds $\alpha \sum |b_j|$. It can shrink some coefficients **exactly to zero**.

Mini Review - Searching for Alpha

REVIEW



- Training, Validation and Testing sets: tune α on the **validation** set.
- A little bit of regularization goes a long way. Ridge was best at $\alpha = 0.9$.

K-Nearest Neighbors

PREDICT

the 3 nearest neighbors

neighbor	class	price
1	A	10
2	A	12
3	B	20

With $k = 3$, what does KNN predict for the new point: the class, and the price?

Hint. Classification takes a vote. Regression takes an average.

K-Nearest Neighbors

ANSWER

With $k = 3$, what does KNN predict: the class and the price?

class: A (2 votes to 1)
price: 14 (10 + 12 + 20) / 3

We use the same three neighbors both times. KNeighborsClassifier takes a vote and KNeighborsRegressor takes the average.

K-Nearest Neighbors

ASK

KNN with $k = 1$:
training accuracy 1.00,
test accuracy 0.82.
What is going on?

K-Nearest Neighbors

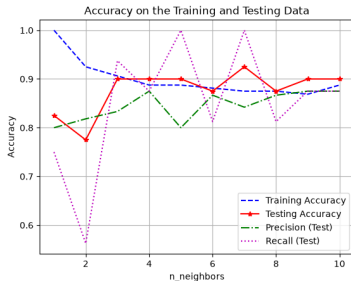
ANSWER

$k = 1$: training accuracy 1.00, test 0.82. What is going on?

- With $k = 1$ every training point is its own nearest neighbor, so the training score is always perfect.
- The model memorized the training data, this is overfitting!
- Try more neighbors and compare training and testing accuracy.

Over and Underfitting

ASK



Where on this plot is the model overfitting? Where would it underfit? Which k would you pick?

Over and Underfitting

ANSWER

Where on this plot is the model overfitting? Where would it underfit? Which k would you pick?

- Small k : training accuracy is much higher than testing accuracy, so the model is overfitting.
- A very big k is too smooth and both scores drop: underfitting.
- Pick a k in the middle where the testing accuracy is good, and think about whether you need precision or recall.

Scale the Data

ASK

One feature is house size (500 to 5000).

Another is number of bedrooms (1 to 5).

Why does KNN need you to scale them?

Scale the Data

ANSWER

House size goes from 500 to 5000 and bedrooms from 1 to 5. Why does KNN need you to scale them?

- KNN uses distance. A difference of 1000 square feet is a much bigger number than a difference of 3 bedrooms, so house size controls the distance.
- Without scaling, the bedrooms barely change the distance, even if they are really important.
- StandardScaler puts every feature on a similar scale: mean 0, standard deviation 1. Fit it on the training data only.

Regularization

ASK

In your own words:
what does regularization do
to the coefficients?
Why does that help with overfitting?

Regularization

ANSWER

In your own words: what does regularization do to the coefficients? Why does that help with overfitting?

- It adds a penalty for big coefficients, so the model keeps them small.
- Big, unstable coefficients usually mean the model is fitting the noise in the training data.
- Smaller coefficients give a simpler model that usually does better on new data.

Regularization

ASK

Ridge or Lasso:

which one can make a coefficient
exactly zero?

When is that useful?

Regularization

ANSWER

Which can make a coefficient exactly zero, Ridge or Lasso? When is it useful?

- Lasso. It can shrink some coefficients exactly to zero, so it removes those features.
- Useful when you think only a few features really matter: it does feature selection for you.
- Ridge keeps every feature and shrinks them all. It helps with many small effects and correlated features.

Regularization

ASK

Why do we choose α
with the validation set
and not the test set?

Regularization

ANSWER

Why do we choose α with the validation set and not the test set?

- Choosing α is part of training the model. If we choose it with the test set, the test set is not a fair test any more.
- We can use the validation set as many times as we want while we pick α .
- You only get to use your test set once, at the very end.

Data Science Interview Practice

INTERVIEW

What are bias and variance
in a machine learning model?
What is the bias-variance trade-off?

Data Science Interview Practice

INTERVIEW

What is the difference between
a parametric and a non-parametric model?
Give an example of each.

Questions?

Questions?