

One-Hot Encoding & K Nearest Neighbors

DATA 3301 Introduction to Data Science Lecture Notes

Name: _____ Date: _____

Key terms (define them in your own words)

categorical variable	
one-hot / dummy variable	
target	
features	
training set	
supervised learning	
K nearest neighbors	

1. Distances need numbers

- **Numeric variable:** a number where _____ makes sense. **Categorical variable:** a _____ from a fixed set of options.
- Euclidean distance _____ values, so every column must be a _____.
- Numbers can still be categorical: zip codes, the 0/1 _____ code in the Parkinson's data.
- Coding Attchd = 1, Detchd = 2, None = 3 invents an _____ and fake _____: None looks _____ as far from Attchd as Detchd does.

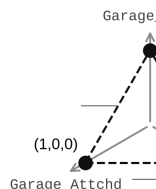
2. One-hot encoding

- Make one new 0/1 column per _____. Each row has exactly _____ 1. These columns are called _____ (or indicator) variables.

Fill in:

Garage Type	Garage_Attchd	Garage_Detchd	Garage_None
Detchd	___	___	___
None	___	___	___
Attchd	___	___	___

- Each level is one step along its own _____.
- Every pair of levels is



- distance apart: $\sqrt{2} \approx$ _____.
- Different garage types add _____ to the squared distance. Same type adds _____.
- K levels $\rightarrow$ _____ columns. House Style (8 levels) $\rightarrow$ _____ columns.

3. One-hot encoding in Python

```
pd.get_dummies(small["Garage Type"], dtype=_____)
# without dtype=int you get _____
```

- On a whole DataFrame, get_dummies only converts _____ columns. prefix sets the start of the new column names.
- OneHotEncoder: .fit() learns the _____. .transform() writes the _____.
- **Sparse** = stores only the _____ entries. Use sparse_output=_____ to see a regular array.
- The result is a NumPy array with no column names. Get them back with enc._____.()
- Use get_dummies to _____. Use OneHotEncoder when building a model: it _____ the levels for new data.

4. Column transformers

```
ct = make_column_transformer(
    (_____, ["Gr Liv Area", ...]),
    (_____, (sparse_output=False),
    ["House Style", "Garage Type"]),
```

- Each argument is a _____ pair. 4 numeric + 8 + 7 dummies = _____ columns.
- remainder="drop": unlisted columns are _____.
"passthrough": copied over _____ (not scaled).

- **Prediction problem:** use observations we _____ have to guess a value we _____ yet.
- Parker predicted 1986 from _____. Ashenfelter predicted it from _____ records, without tasting.
- Prices are known for 1952 to _____ and unknown for 1981 to 1991. Price is relative: 1961 = _____.

- **Target** = the column we _____. **Features** = the columns we predict _____ (also called _____ variables).
- **Training set** = rows where the target is _____. **New observation** = target unknown, like _____.
- **Supervised learning**: learn from examples where the _____ is known.
Regression: target is a _____. **Classification**: target is a _____.
- **T / F**: ____ Predicting "good" or "bad" vintage is regression.
- We predict $\log(\text{price})$. $\text{np.log}()$ turns each doubling into adding _____.
 $\text{np.} ______ ()$ undoes it.

A scatter plot showing the relationship between summer temperature (std.) on the y-axis and winter rain (std.) on the x-axis. The y-axis ranges from -1.0 to 0.5, and the x-axis ranges from -1.5 to 1.0. A single data point for the year 1986 is highlighted with a black star at approximately (-0.4, -0.2). Other data points are represented by open circles.

- **Circle** the 5 nearest neighbors of 1986.
- 1. Pick a number ____.
- 2. Find the K _____ training rows (_____ distance).
- 3. _____ their targets.
- 1986's neighbors: 2.40, 2.42, 2.48, 2.50, 2.52 → prediction = _____.

- ## 8. KNN in scikit-learn

```
model = _____(n_neighbors=_____)
model._____(X=X_train, y=bordeaux_train["log(price)"])
model._____(X=X_new)
# scale new data with the training scaler:
X_new = scaler._____(bordeaux_test[features])
```

- ## 9. Both halves together

- 1,500 sq ft, 1Story, attached garage: \$215,434. Same house with a detached garage: \$_____. All 5 neighbors changed to _____ houses.
- Why? A garage mismatch adds _____ to the squared distance, more than a few hundred sq ft once area is _____.
- **Exit question:** why must Garage Type be one-hot encoded before KNN can use it?
