Mudcard

- **I was struggling to determine whether a variable was continuous or ordinal such as age or gross pay**
    - Yeah, there is no easy way to give definite answer to such questions and the answer might depend on the dataset and problem you are trying to solve
    - If you really can't decide, develop two identical ML pipelines with one difference: treat age as continuous in one model; and as ordinal in the other model. Compare the validation scores of the two model and choos the one that gives you a better score!
    - This is called data-driven decision making. You perform an experiment to decide which approach is better.
- **I'm just asking about the visualization application. In the last part, you shared a link of https://pyviz.org/, where I saw the application of the dashboard, which is extremely useful (I was trying to build one during the internship). Is there any chance that we can cover this part or any other part of reporting our project during the semester?**
    - Ufortunately we don't have time to cover visualizations in more detail but check out DATA1500 during the spring term. The whole course is about visualizations!
- **Will visualizing data from multiple columns better demonstrate technical strength for the final project?**
    - If you plan to show scatter matrices in your final report/presentation, then the answer is nope. It will demonatrate that you have no clear idea what you want to visuzalize :)
        - Your figure should have a goal, a message you are trying to convey to your audience.
        - Scatter matrices are extremely messy and bad for this purpose.
    - if you plan to create 3D plots, I'd advise you against it. those are pretty difficult to make sense of unless they are interactive.
- **When should we do log-transformations to bins and how exactly should we do it?**
    - How exactly you should do it is in the lecture notes, check out the code above the figure with the log bins
    - You should use log transformation if the quantity you want to visualize varies several orders of magnitudes. For example, salaries can be on the order of 10k USD to well above 10^6 - 10^7 USD which is 2-3 orders of magnitude.
- **Processing the data before creating visualizations is confusing at times. Sometimes we need to create a matrix, sometimes we need to get counts, other times we just need a column. It's a bit confusing which cases we need when and why.**
    - Not sure I understand the question. We didn't process the data yet. We just visualized 1, 2, or a couple of columns so far.
    - Come to the office hours and talk to me or the TAs, or post on the course forum.
- **The heatmap visualization, even when refined, feels oddly vague and hard to conceptualize. Could you show an example of when a heatmap is clearly the optimal visualization tool?**
    - It is the optimal visualiation tool when a scatter plot fails for some reason (usually because there are too many overlaping points on the scatter plot).
- **for the second quiz in this lecture, where do we go to find the variable type (categorical vs. continuous), i think you mentioned finding it in github but I'm not sure where to go!**
    - the course's github repo has a text file, the dataset description
    - but you can also just do .descibe() or .value_counts() in the notebook to quickly figure out the feature properties.
- **In the case of one catergorical data type, when should you use a histogram and when should you use a bar plot? Or is it essentially the same and its just data-dependent?**
    - you always use a bar plot with categorical data, and histograms with continuous data
    - the main difference is that the order of the bars does not matter when you visualize categorical data; but the order of the histogram bars matter a lot! The bins are ordered!

Split iid data

By the end of this lecture, you will be able to

- apply basic split to iid datasets
- apply k-fold split to iid datasets
- apply stratified splits to imbalanced data

The supervised ML pipeline

The goal: Use the training data (X and y) to develop a model which can accurately predict the target variable (y_new') for previously unseen data (X_new).

1. Exploratory Data Analysis (EDA): you need to understand your data and verify that it doesn't contain errors

- do as much EDA as you can!

2. Split the data into different sets: most often the sets are train, validation, and test (or holdout)

- practitioners often make errors in this step!
- you can split the data randomly, based on groups, based on time, or any other non-standard way if necessary to answer your ML question

3. Preprocess the data: ML models only work if X and Y are numbers! Some ML models additionally require each feature to have 0 mean and 1 standard deviation (standardized features)

- often the original features you get contain strings (for example a gender feature would contain 'male', 'female', 'non-binary', 'unknown') which needs to transformed into numbers
- often the features are not standardized (e.g., age is between 0 and 100) but it needs to be standardized

4. Choose an evaluation metric: depends on the priorities of the stakeholders

- often requires quite a bit of thinking and ethical considerations

5. Choose one or more ML techniques: it is highly recommended that you try multiple models

- start with simple models like linear or logistic regression
- try also more complex models like nearest neighbors, support vector machines, random forest, etc.

6. Tune the hyperparameters of your ML models (aka cross-validation)

- ML techniques have hyperparameters that you need to optimize to achieve best performance
- for each ML model, decide which parameters to tune and what values to try
- loop through each parameter combination
    - train one model for each parameter combination
    - evaluate how well the model performs on the validation set
- take the parameter combo that gives the best validation score
- evaluate that model on the test set to report how well the model is expected to perform on previously unseen data

7. Interpret your model: black boxes are often not useful

- check if your model uses features that make sense (excellent tool for debugging)
- often model predictions are not enough, you need to be able to explain how the model arrived to a particular prediction (e.g., in health care)

Why do we split the data?

- we want to find the best hyper-parameters of our ML algorithms
    - fit models to training data
    - evaluate each model on validation set
    - we find hyper-parameter values that optimize the validation score
- we want to know how the model will perform on previously unseen data
    - apply our final model on the test set

We need to split the data into three parts!

Recap from the second lecture

- **the learner's input**
    - Domain set $\mathcal{X}$ - a set of objects we wish to label.
    - Label set $\mathcal{Y}$ - a set of possible labels.
    - Training data $S = ((x_1, y_1), \ldots, (x_m, y_m))$ - a finite sequence of pairs from $\mathcal{X}$, $\mathcal{Y}$. This is what the learner has access to.
        - $X = (x_1, \ldots, x_m)$ is the feature matrix which is usually a 2D matrix, and $Y = (y_1, \ldots, y_m)$ is the target variable which is a vector.
- let's denote the probability distribution over $\mathcal{X}$ by D.
- let's assume there is some correct labeling function $f : \mathcal{X} \to \mathcal{Y}$.
- **a training example is then generated by sampling x_i from D, and the label y_i is generated using f.**

I.I.D. assumption

- **the i.i.d. assumption**: the examples in the training set are independently and identically distributed according to D
  - every x_i is freshly sampled from D and then labelled by f
  - that is, x_i and y_i are picked independently of the other instances
  - S is a window through which the learner gets partial info about D and the labeling function f
  - the larger the sample gets, the more likely it is to reflect more accurately D and f
- examples of not iid data:
  - data generated by time-dependent processes
  - data has group structure (samples collected from e.g., different subjects, experiments, measurement devices)

Split iid data

By the end of this lecture, you will be able to

- **apply basic split to iid datasets**
- apply k-fold split to iid datasets
- apply stratified splits to imbalanced data

Splitting strategies for iid data: basic approach

- 60% train, 20% validation, 20% test for small datasets
- 98% train, 1% validation, 1% test for large datasets
  - if you have 1 million points, you still have 10000 points in validation and test which is plenty to assess model performance

Let's work with the adult data!

In [1]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split

df = pd.read_csv('data/adult_test.csv')

# let's separate the feature matrix X, and target variable y
y = df['gross-income'] # remember, we want to predict who earns more than 50k or less than 50k
X = df.loc[:, df.columns != 'gross-income'] # all other columns are features
print(y)
print(X.head())
```

```
0         <=50K.
1         <=50K.
2          >50K.
3          >50K.
4         <=50K.
          ...
16276     <=50K.
16277     <=50K.
16278     <=50K.
16279     <=50K.
16280      >50K.
Name: gross-income, Length: 16281, dtype: object
   age   workclass  fnlwgt      education  education-num       marital-status  \
0   25     Private  226802           11th              7        Never-married
1   38     Private   89814        HS-grad              9   Married-civ-spouse
2   28   Local-gov  336951     Assoc-acdm             12   Married-civ-spouse
3   44     Private  160323   Some-college             10   Married-civ-spouse
4   18           ?  103497   Some-college             10        Never-married

          occupation relationship   race      sex  capital-gain  \
0  Machine-op-inspct    Own-child  Black     Male             0
1    Farming-fishing      Husband  White     Male             0
2    Protective-serv      Husband  White     Male             0
3  Machine-op-inspct      Husband  Black     Male          7688
4                  ?    Own-child  White   Female             0

   capital-loss  hours-per-week  native-country
0             0              40   United-States
1             0              50   United-States
2             0              40   United-States
3             0              40   United-States
4             0              30   United-States
```

In [2]:
```
help(train_test_split)
```

```
Help on function train_test_split in module sklearn.model_selection._split:

train_test_split(*arrays, test_size=None, train_size=None, random_state=None, shuffle=True, stratify=None)
    Split arrays or matrices into random train and test subsets.

    Quick utility that wraps input validation,
    ``next(ShuffleSplit().split(X, y))``, and application to input data
    into a single call for splitting (and optionally subsampling) data into a
    one-liner.

    Read more in the :ref:`User Guide <cross_validation>`.

    Parameters
    ----------
    *arrays : sequence of indexables with same length / shape[0]
        Allowed inputs are lists, numpy arrays, scipy-sparse
        matrices or pandas dataframes.

    test_size : float or int, default=None
        If float, should be between 0.0 and 1.0 and represent the proportion
        of the dataset to include in the test split. If int, represents the
        absolute number of test samples. If None, the value is set to the
        complement of the train size. If ``train_size`` is also None, it will
        be set to 0.25.

    train_size : float or int, default=None
        If float, should be between 0.0 and 1.0 and represent the
        proportion of the dataset to include in the train split. If
        int, represents the absolute number of train samples. If None,
        the value is automatically set to the complement of the test size.

    random_state : int, RandomState instance or None, default=None
        Controls the shuffling applied to the data before applying the split.
        Pass an int for reproducible output across multiple function calls.
        See :term:`Glossary <random_state>`.

    shuffle : bool, default=True
        Whether or not to shuffle the data before splitting. If shuffle=False
        then stratify must be None.

    stratify : array-like, default=None
        If not None, data is split in a stratified fashion, using this as
        the class labels.
        Read more in the :ref:`User Guide <stratification>`.

    Returns
    -------
    splitting : list, length=2 * len(arrays)
        List containing train-test split of inputs.

        .. versionadded:: 0.16
            If the input is sparse, the output will be a
            ``scipy.sparse.csr_matrix``. Else, output type is the same as the
            input type.

    Examples
    --------
    >>> import numpy as np
    >>> from sklearn.model_selection import train_test_split
    >>> X, y = np.arange(10).reshape((5, 2)), range(5)
    >>> X
    array([[0, 1],
           [2, 3],
           [4, 5],
           [6, 7],
           [8, 9]])
    >>> list(y)
    [0, 1, 2, 3, 4]

    >>> X_train, X_test, y_train, y_test = train_test_split(
    ...     X, y, test_size=0.33, random_state=42)
    ...
    >>> X_train
    array([[4, 5],
           [0, 1],
           [6, 7]])
    >>> y_train
    [2, 0, 3]
    >>> X_test
    array([[2, 3],
           [8, 9]])
    >>> y_test
    [1, 4]

    >>> train_test_split(y, shuffle=False)
    [[0, 1, 2], [3, 4]]
```

```
In [3]: random_state = 42

# first split to separate out the training set
X_train, X_other, y_train, y_other = train_test_split(X,y,\
                    train_size = 0.6,random_state = random_state)
print('training set:',X_train.shape, y_train.shape) # 60% of points are in train
print(X_other.shape, y_other.shape) # 40% of points are in other

# second split to separate out the validation and test sets
X_val, X_test, y_val, y_test = train_test_split(X_other,y_other,\
                    train_size = 0.5,random_state = random_state)
print('validation set:',X_val.shape, y_val.shape) # 20% of points are in validation
print('test set:',X_test.shape, y_test.shape) # 20% of points are in test

print(X_train.head())
```

```
training set: (9768, 14) (9768,)
(6513, 14) (6513,)
validation set: (3256, 14) (3256,)
test set: (3257, 14) (3257,)
       age          workclass  fnlwgt      education  education-num  \
4050    22            Private  335950        HS-grad              9
11446   29            Private   78261        HS-grad              9
12427   74   Self-emp-not-inc  160009     Assoc-acdm             12
5702    39       Self-emp-inc   31709   Some-college             10
13058   50            Private  144084        HS-grad              9

           marital-status        occupation   relationship    race      sex  \
4050         Never-married     Other-service  Not-in-family   Black     Male
11446            Separated   Protective-serv  Not-in-family   White     Male
12427   Married-civ-spouse   Exec-managerial        Husband   White     Male
5702    Married-civ-spouse      Adm-clerical           Wife   White   Female
13058            Separated             Sales      Unmarried   White   Female

       capital-gain  capital-loss  hours-per-week  native-country
4050              0             0              70   United-States
11446             0             0              55   United-States
12427             0             0              30   United-States
5702              0             0              20   United-States
13058             0             0              55   United-States
```

Randomness due to splitting

- the model performance, validation and test scores will change depending on which points are in train, val, test
  - inherent randomness or uncertainty of the ML pipeline
- change the random state a couple of times and repeat the whole ML pipeline to assess how much the random splitting affects your test score
  - you would expect a similar uncertainty when the model is deployed

Quiz 1

What's the second train_test_split line if you want to end up with 60-20-20 in train-val-test? Print out the sizes of X_train, X_val, X_test to verify!

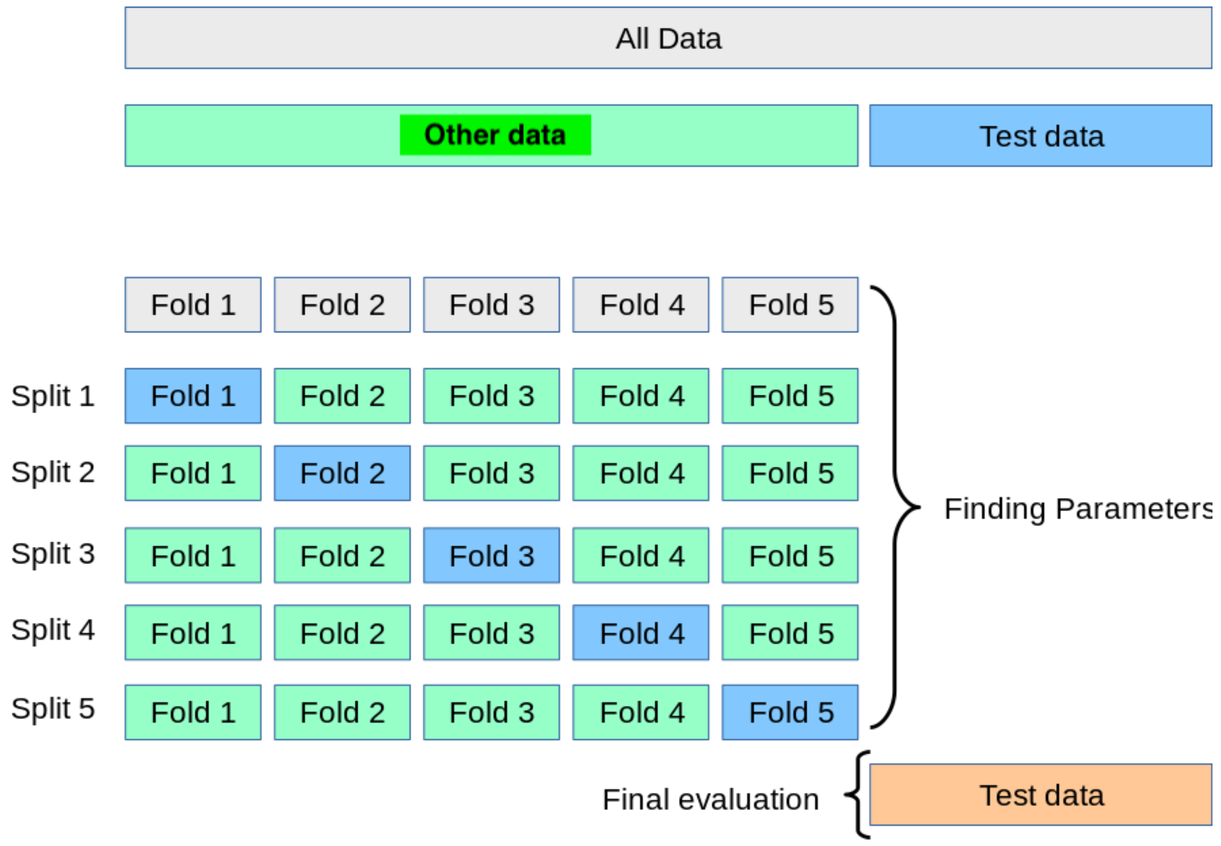
```
In [4]: X_other, X_test, y_other, y_test = train_test_split(X,y,\
                    train_size = 0.8,random_state=random_state)
# add your line below and choose the correct solution from canvas
```

Split iid data

By the end of this lecture, you will be able to

- apply basic split to iid datasets
- **apply k-fold split to iid datasets**
- apply stratified splits to imbalanced data

Other splitting strategy for iid data: k-fold splitting



In [5]:
```
from sklearn.model_selection import KFold
help(KFold)
```

```
Help on class KFold in module sklearn.model_selection._split:

class KFold(_UnsupportedGroupCVMixin, _BaseKFold)
 |  KFold(n_splits=5, *, shuffle=False, random_state=None)
 |
 |  K-Fold cross-validator.
 |
 |  Provides train/test indices to split data in train/test sets. Split
 |  dataset into k consecutive folds (without shuffling by default).
 |
 |  Each fold is then used once as a validation while the k - 1 remaining
 |  folds form the training set.
 |
 |  Read more in the :ref:`User Guide <k_fold>`.
 |
 |  For visualisation of cross-validation behaviour and
 |  comparison between common scikit-learn split methods
 |  refer to :ref:`sphx_glr_auto_examples_model_selection_plot_cv_indices.py`
 |
 |  Parameters
 |  ----------
 |  n_splits : int, default=5
 |      Number of folds. Must be at least 2.
 |
 |      .. versionchanged:: 0.22
 |          ``n_splits`` default value changed from 3 to 5.
 |
 |  shuffle : bool, default=False
 |      Whether to shuffle the data before splitting into batches.
 |      Note that the samples within each split will not be shuffled.
 |
 |  random_state : int, RandomState instance or None, default=None
 |      When `shuffle` is True, `random_state` affects the ordering of the
 |      indices, which controls the randomness of each fold. Otherwise, this
 |      parameter has no effect.
 |      Pass an int for reproducible output across multiple function calls.
 |      See :term:`Glossary <random_state>`.
 |
 |  Examples
 |  --------
 |  >>> import numpy as np
 |  >>> from sklearn.model_selection import KFold
 |  >>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
 |  >>> y = np.array([1, 2, 3, 4])
 |  >>> kf = KFold(n_splits=2)
 |  >>> kf.get_n_splits(X)
 |  2
 |  >>> print(kf)
 |  KFold(n_splits=2, random_state=None, shuffle=False)
 |  >>> for i, (train_index, test_index) in enumerate(kf.split(X)):
 |  ...     print(f"Fold {i}:")
 |  ...     print(f"  Train: index={train_index}")
 |  ...     print(f"  Test:  index={test_index}")
 |  Fold 0:
 |    Train: index=[2 3]
 |    Test:  index=[0 1]
 |  Fold 1:
 |    Train: index=[0 1]
 |    Test:  index=[2 3]
 |
 |  Notes
 |  -----
 |  The first ``n_samples % n_splits`` folds have size
 |  ``n_samples // n_splits + 1``, other folds have size
 |  ``n_samples // n_splits``, where ``n_samples`` is the number of samples.
 |
 |  Randomized CV splitters may return different results for each call of
 |  split. You can make the results identical by setting `random_state`
 |  to an integer.
 |
 |  See Also
 |  --------
 |  StratifiedKFold : Takes class information into account to avoid building
 |      folds with imbalanced class distributions (for binary or multiclass
 |      classification tasks).
 |
 |  GroupKFold : K-fold iterator variant with non-overlapping groups.
 |
 |  RepeatedKFold : Repeats K-Fold n times.
 |
 |  Method resolution order:
 |      KFold
 |      _UnsupportedGroupCVMixin
 |      _BaseKFold
 |      BaseCrossValidator
 |      sklearn.utils._metadata_requests._MetadataRequester
 |      builtins.object
```

```
 |
 |  Methods defined here:
 |
 |  __init__(self, n_splits=5, *, shuffle=False, random_state=None)
 |      Initialize self.  See help(type(self)) for accurate signature.
 |
 |  ----------------------------------------------------------------------
 |  Data and other attributes defined here:
 |
 |  __abstractmethods__ = frozenset()
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from _UnsupportedGroupCVMixin:
 |
 |  split(self, X, y=None, groups=None)
 |      Generate indices to split data into training and test set.
 |
 |      Parameters
 |      ----------
 |      X : array-like of shape (n_samples, n_features)
 |          Training data, where `n_samples` is the number of samples
 |          and `n_features` is the number of features.
 |
 |      y : array-like of shape (n_samples,)
 |          The target variable for supervised learning problems.
 |
 |      groups : object
 |          Always ignored, exists for compatibility.
 |
 |      Yields
 |      ------
 |      train : ndarray
 |          The training set indices for that split.
 |
 |      test : ndarray
 |          The testing set indices for that split.
 |
 |  ----------------------------------------------------------------------
 |  Data descriptors inherited from _UnsupportedGroupCVMixin:
 |
 |  __dict__
 |      dictionary for instance variables
 |
 |  __weakref__
 |      list of weak references to the object
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from _BaseKFold:
 |
 |  get_n_splits(self, X=None, y=None, groups=None)
 |      Returns the number of splitting iterations in the cross-validator.
 |
 |      Parameters
 |      ----------
 |      X : object
 |          Always ignored, exists for compatibility.
 |
 |      y : object
 |          Always ignored, exists for compatibility.
 |
 |      groups : object
 |          Always ignored, exists for compatibility.
 |
 |      Returns
 |      -------
 |      n_splits : int
 |          Returns the number of splitting iterations in the cross-validator.
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from BaseCrossValidator:
 |
 |  __repr__(self)
 |      Return repr(self).
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from sklearn.utils._metadata_requests._MetadataRequester:
 |
 |  get_metadata_routing(self)
 |      Get metadata routing of this object.
 |
 |      Please check :ref:`User Guide <metadata_routing>` on how the routing
 |      mechanism works.
 |
 |      Returns
 |      -------
 |      routing : MetadataRequest
 |          A :class:`~sklearn.utils.metadata_routing.MetadataRequest` encapsulating
```

```
 |          routing information.
 |
 |  ----------------------------------------------------------------------
 |  Class methods inherited from sklearn.utils._metadata_requests._MetadataRequester:
 |
 |  __init_subclass__(**kwargs)
 |      Set the ``set_{method}_request`` methods.
 |
 |      This uses PEP-487 [1]_ to set the ``set_{method}_request`` methods. It
 |      looks for the information available in the set default values which are
 |      set using ``__metadata_request__*`` class attributes, or inferred
 |      from method signatures.
 |
 |      The ``__metadata_request__*`` class attributes are used when a method
 |      does not explicitly accept a metadata through its arguments or if the
 |      developer would like to specify a request value for those metadata
 |      which are different from the default ``None``.
 |
 |      References
 |      ----------
 |      .. [1] https://www.python.org/dev/peps/pep-0487
```

In [6]:

```
random_state =42

# first split to separate out the test set
X_other, X_test, y_other, y_test = train_test_split(X,y,test_size = 0.2,random_state=random_state)
print(X_other.shape,y_other.shape)
print('test set:',X_test.shape,y_test.shape)

# do KFold split on other
kf = KFold(n_splits=5,shuffle=True,random_state=random_state)
for train_index, val_index in kf.split(X_other,y_other):
    X_train = X_other.iloc[train_index]
    y_train = y_other.iloc[train_index]
    X_val = X_other.iloc[val_index]
    y_val = y_other.iloc[val_index]
    print('   training set:',X_train.shape, y_train.shape)
    print('   validation set:',X_val.shape, y_val.shape)
    # the validation set contains different points in each iteration
    print(X_val[['age','workclass','education']].head())
```

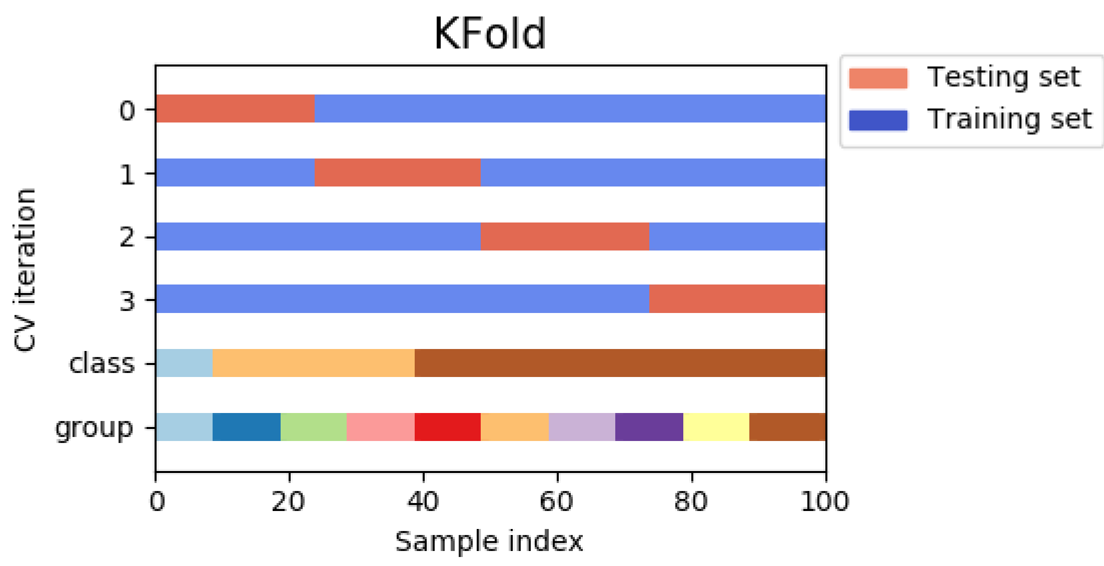
```
(13024, 14) (13024,)
test set: (3257, 14) (3257,)
   training set: (10419, 14) (10419,)
   validation set: (2605, 14) (2605,)
       age          workclass      education
9850    59            Private   Some-college
103     58   Self-emp-not-inc            9th
1383    45            Private        HS-grad
11034   49   Self-emp-not-inc      Bachelors
14876   59   Self-emp-not-inc      Bachelors
   training set: (10419, 14) (10419,)
   validation set: (2605, 14) (2605,)
       age     workclass      education
13384   60   Federal-gov      Bachelors
8471    20       Private        HS-grad
13406   21             ?   Some-college
13394   35       Private        HS-grad
15123   38       Private   Some-college
   training set: (10419, 14) (10419,)
   validation set: (2605, 14) (2605,)
       age     workclass      education
647     60             ?      Bachelors
9314    26       Private   Some-college
14499   52       Private        HS-grad
7332    53   Federal-gov     Assoc-acdm
12523   21       Private           10th
   training set: (10419, 14) (10419,)
   validation set: (2605, 14) (2605,)
       age workclass      education
5294    53   Private        HS-grad
3481    41   Private        HS-grad
7671    49   Private   Some-college
11055   39   Private      Bachelors
12751   18         ?           12th
   training set: (10420, 14) (10420,)
   validation set: (2604, 14) (2604,)
       age      workclass      education
4265    23              ?           10th
5290    23        Private        HS-grad
1157    56   Self-emp-inc    Prof-school
12344   18        Private           11th
13683   55        Private        HS-grad
```

How many splits should I create?

- tough question, 3-5 is most common
- if you do n splits, n models will be trained, so the larger the n, the most computationally intensive it will be to train the models
- KFold is usually better suited to small datasets
- KFold is good to estimate uncertainty due to random splitting of train and val, but it is not perfect
    - the test set remains the same

Why shuffling iid data is important?

- by default, data is not shuffled by Kfold which can introduce errors!



Quiz 2

Given the labels below, what are the balances of each class?

y = [0,0,0,2,2,0,0,2,0,1]

Split iid data

By the end of this lecture, you will be able to

- apply basic split to iid datasets
- apply k-fold split to iid datasets
- **apply stratified splits to imbalanced data**

Imbalanced data

- imbalanced data: only a small fraction of the points are in one of the classes, usually ~5% or less but there is no hard limit here
- examples:
    - people visit a bank's website. do they sign up for a new credit card?
        - most customers just browse and leave the page
        - usually 1% or less of the customers get a credit card (class 1), the rest leaves the page without signing up (class 0).
    - fraud detection
        - only a tiny fraction of credit card payments are fraudulent
    - rare disease diagnosis
- the issue with imbalanced data:
    - if you apply train_test_split or KFold, you might not have class 1 points in one of your sets by chance
    - this is what we need to fix

Solution: stratified splits

In [7]:
```
df = pd.read_csv('data/imbalanced_data.csv')

X = df[['feature1','feature2']]
y = df['y']

print(y.value_counts())
```

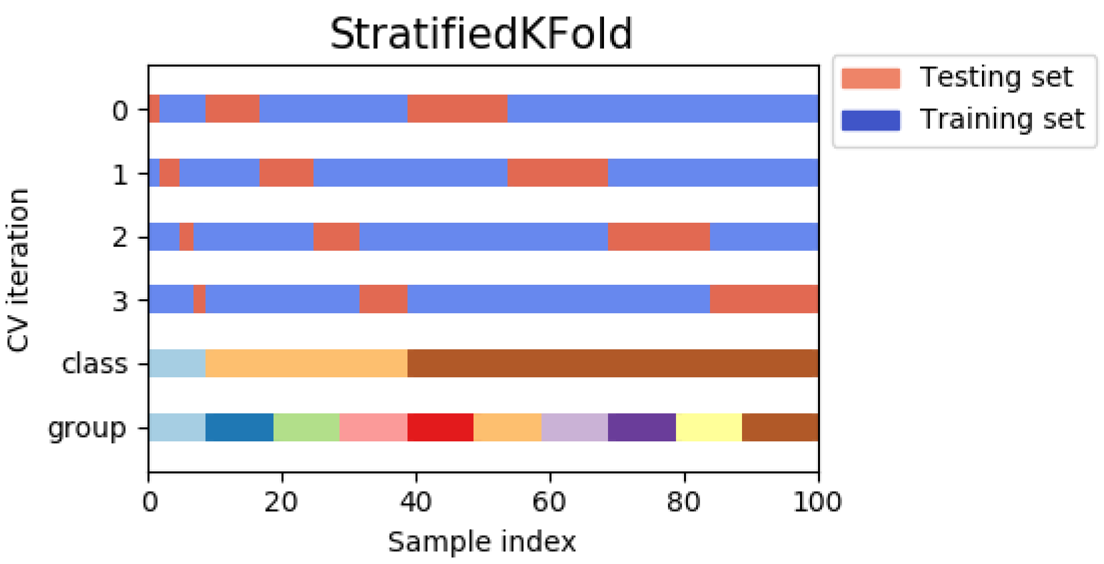
```
y
0    990
1     10
Name: count, dtype: int64
```

In [19]:
```
# 4 and 10
```

```python
random_state = 42

X_train, X_other, y_train, y_other = train_test_split(X,y,train_size = 0.6,random_state=random_state)
X_val, X_test, y_val, y_test = train_test_split(X_other,y_other,train_size = 0.5,random_state=random_state)

print('**balance without stratification:**')
# a variation on the order of 1% which would be too much for imbalanced data!
print(np.unique(y_train,return_counts=True))
print(np.unique(y_val,return_counts=True))
print(np.unique(y_test,return_counts=True))

X_train, X_other, y_train, y_other = train_test_split(X,y,train_size = 0.6,stratify=y,random_state=random_state)
X_val, X_test, y_val, y_test = train_test_split(X_other,y_other,train_size = 0.5,stratify=y_other,random_state=rand
print('**balance with stratification:**')
# very little variation (in the 4th decimal point only) which is important if the problem is imbalanced
print(np.unique(y_train,return_counts=True))
print(np.unique(y_val,return_counts=True))
print(np.unique(y_test,return_counts=True))
```

```
**balance without stratification:**
(array([0, 1]), array([597,   3]))
(array([0, 1]), array([197,   3]))
(array([0, 1]), array([196,   4]))
**balance with stratification:**
(array([0, 1]), array([594,   6]))
(array([0, 1]), array([198,   2]))
(array([0, 1]), array([198,   2]))
```

Stratified folds



In [9]:
```python
from sklearn.model_selection import StratifiedKFold
help(StratifiedKFold)
```

```
Help on class StratifiedKFold in module sklearn.model_selection._split:

class StratifiedKFold(_BaseKFold)
 |  StratifiedKFold(n_splits=5, *, shuffle=False, random_state=None)
 |
 |  Stratified K-Fold cross-validator.
 |
 |  Provides train/test indices to split data in train/test sets.
 |
 |  This cross-validation object is a variation of KFold that returns
 |  stratified folds. The folds are made by preserving the percentage of
 |  samples for each class.
 |
 |  Read more in the :ref:`User Guide <stratified_k_fold>`.
 |
 |  For visualisation of cross-validation behaviour and
 |  comparison between common scikit-learn split methods
 |  refer to :ref:`sphx_glr_auto_examples_model_selection_plot_cv_indices.py`
 |
 |  Parameters
 |  ----------
 |  n_splits : int, default=5
 |      Number of folds. Must be at least 2.
 |
 |      .. versionchanged:: 0.22
 |          ``n_splits`` default value changed from 3 to 5.
 |
 |  shuffle : bool, default=False
 |      Whether to shuffle each class's samples before splitting into batches.
 |      Note that the samples within each split will not be shuffled.
 |
 |  random_state : int, RandomState instance or None, default=None
 |      When `shuffle` is True, `random_state` affects the ordering of the
 |      indices, which controls the randomness of each fold for each class.
 |      Otherwise, leave `random_state` as `None`.
 |      Pass an int for reproducible output across multiple function calls.
 |      See :term:`Glossary <random_state>`.
 |
 |  Examples
 |  --------
 |  >>> import numpy as np
 |  >>> from sklearn.model_selection import StratifiedKFold
 |  >>> X = np.array([[1, 2], [3, 4], [1, 2], [3, 4]])
 |  >>> y = np.array([0, 0, 1, 1])
 |  >>> skf = StratifiedKFold(n_splits=2)
 |  >>> skf.get_n_splits(X, y)
 |  2
 |  >>> print(skf)
 |  StratifiedKFold(n_splits=2, random_state=None, shuffle=False)
 |  >>> for i, (train_index, test_index) in enumerate(skf.split(X, y)):
 |  ...     print(f"Fold {i}:")
 |  ...     print(f"  Train: index={train_index}")
 |  ...     print(f"  Test:  index={test_index}")
 |  Fold 0:
 |    Train: index=[1 3]
 |    Test:  index=[0 2]
 |  Fold 1:
 |    Train: index=[0 2]
 |    Test:  index=[1 3]
 |
 |  Notes
 |  -----
 |  The implementation is designed to:
 |
 |  * Generate test sets such that all contain the same distribution of
 |    classes, or as close as possible.
 |  * Be invariant to class label: relabelling ``y = ["Happy", "Sad"]`` to
 |    ``y = [1, 0]`` should not change the indices generated.
 |  * Preserve order dependencies in the dataset ordering, when
 |    ``shuffle=False``: all samples from class k in some test set were
 |    contiguous in y, or separated in y by samples from classes other than k.
 |  * Generate test sets where the smallest and largest differ by at most one
 |    sample.
 |
 |  .. versionchanged:: 0.22
 |      The previous implementation did not follow the last constraint.
 |
 |  See Also
 |  --------
 |  RepeatedStratifiedKFold : Repeats Stratified K-Fold n times.
 |
 |  Method resolution order:
 |      StratifiedKFold
 |      _BaseKFold
 |      BaseCrossValidator
 |      sklearn.utils._metadata_requests._MetadataRequester
 |      builtins.object
```

```
 |
 |  Methods defined here:
 |
 |  __init__(self, n_splits=5, *, shuffle=False, random_state=None)
 |      Initialize self.  See help(type(self)) for accurate signature.
 |
 |  split(self, X, y, groups=None)
 |      Generate indices to split data into training and test set.
 |
 |      Parameters
 |      ----------
 |      X : array-like of shape (n_samples, n_features)
 |          Training data, where `n_samples` is the number of samples
 |          and `n_features` is the number of features.
 |
 |          Note that providing ``y`` is sufficient to generate the splits and
 |          hence ``np.zeros(n_samples)`` may be used as a placeholder for
 |          ``X`` instead of actual training data.
 |
 |      y : array-like of shape (n_samples,)
 |          The target variable for supervised learning problems.
 |          Stratification is done based on the y labels.
 |
 |      groups : object
 |          Always ignored, exists for compatibility.
 |
 |      Yields
 |      ------
 |      train : ndarray
 |          The training set indices for that split.
 |
 |      test : ndarray
 |          The testing set indices for that split.
 |
 |      Notes
 |      -----
 |      Randomized CV splitters may return different results for each call of
 |      split. You can make the results identical by setting `random_state`
 |      to an integer.
 |
 |  ----------------------------------------------------------------------
 |  Data and other attributes defined here:
 |
 |  __abstractmethods__ = frozenset()
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from _BaseKFold:
 |
 |  get_n_splits(self, X=None, y=None, groups=None)
 |      Returns the number of splitting iterations in the cross-validator.
 |
 |      Parameters
 |      ----------
 |      X : object
 |          Always ignored, exists for compatibility.
 |
 |      y : object
 |          Always ignored, exists for compatibility.
 |
 |      groups : object
 |          Always ignored, exists for compatibility.
 |
 |      Returns
 |      -------
 |      n_splits : int
 |          Returns the number of splitting iterations in the cross-validator.
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from BaseCrossValidator:
 |
 |  __repr__(self)
 |      Return repr(self).
 |
 |  ----------------------------------------------------------------------
 |  Methods inherited from sklearn.utils._metadata_requests._MetadataRequester:
 |
 |  get_metadata_routing(self)
 |      Get metadata routing of this object.
 |
 |      Please check :ref:`User Guide <metadata_routing>` on how the routing
 |      mechanism works.
 |
 |      Returns
 |      -------
 |      routing : MetadataRequest
 |          A :class:`~sklearn.utils.metadata_routing.MetadataRequest` encapsulating
 |          routing information.
```

```
|
|  ----------------------------------------------------------------------
|  Class methods inherited from sklearn.utils._metadata_requests._MetadataRequester:
|
|  __init_subclass__(**kwargs)
|      Set the ``set_{method}_request`` methods.
|
|      This uses PEP-487 [1]_ to set the ``set_{method}_request`` methods. It
|      looks for the information available in the set default values which are
|      set using ``__metadata_request__*`` class attributes, or inferred
|      from method signatures.
|
|      The ``__metadata_request__*`` class attributes are used when a method
|      does not explicitly accept a metadata through its arguments or if the
|      developer would like to specify a request value for those metadata
|      which are different from the default ``None``.
|
|      References
|      ----------
|      .. [1] https://www.python.org/dev/peps/pep-0487
|
|  ----------------------------------------------------------------------
|  Data descriptors inherited from sklearn.utils._metadata_requests._MetadataRequester:
|
|  __dict__
|      dictionary for instance variables
|
|  __weakref__
|      list of weak references to the object
```

In [10]:

```
# what we did before: variance in balance on the order of 1%
random_state = 2

X_other, X_test, y_other, y_test = train_test_split(X,y,test_size = 0.2,random_state=random_state)
print('test balance:',np.unique(y_test,return_counts=True))

# do KFold split on other
kf = KFold(n_splits=4,shuffle=True,random_state=random_state)
for train_index, val_index in kf.split(X_other,y_other):
    print('new fold')
    X_train = X_other.iloc[train_index]
    y_train = y_other.iloc[train_index]
    X_val = X_other.iloc[val_index]
    y_val = y_other.iloc[val_index]
    print(np.unique(y_train,return_counts=True))
    print(np.unique(y_val,return_counts=True))
```

```
test balance: (array([0, 1]), array([198,   2]))
new fold
(array([0, 1]), array([596,   4]))
(array([0, 1]), array([196,   4]))
new fold
(array([0, 1]), array([593,   7]))
(array([0, 1]), array([199,   1]))
new fold
(array([0, 1]), array([592,   8]))
(array([0]), array([200]))
new fold
(array([0, 1]), array([595,   5]))
(array([0, 1]), array([197,   3]))
```

In [11]:

```
# stratified K Fold: variation in balance is very small (4th decimal point)
random_state = 42

# stratified train-test split
X_other, X_test, y_other, y_test = train_test_split(X,y,test_size = 0.2,stratify=y,random_state=random_state)
print('test balance:',np.unique(y_test,return_counts=True))

# do StratifiedKFold split on other
kf = StratifiedKFold(n_splits=4,shuffle=True,random_state=random_state)
for train_index, val_index in kf.split(X_other,y_other):
    print('new fold')
    X_train = X_other.iloc[train_index]
    y_train = y_other.iloc[train_index]
    X_val = X_other.iloc[val_index]
    y_val = y_other.iloc[val_index]
    print(np.unique(y_train,return_counts=True))
    print(np.unique(y_val,return_counts=True))
```

```
test balance: (array([0, 1]), array([198,   2]))
new fold
(array([0, 1]), array([594,   6]))
(array([0, 1]), array([198,   2]))
new fold
(array([0, 1]), array([594,   6]))
(array([0, 1]), array([198,   2]))
new fold
(array([0, 1]), array([594,   6]))
(array([0, 1]), array([198,   2]))
new fold
(array([0, 1]), array([594,   6]))
(array([0, 1]), array([198,   2]))
```

Mudcard

In [ ]: