UCI Datasets

CSE 4311 – Machine Learning
Vassilis Athitsos
Computer Science and Engineering Department
University of Texas at Arlington

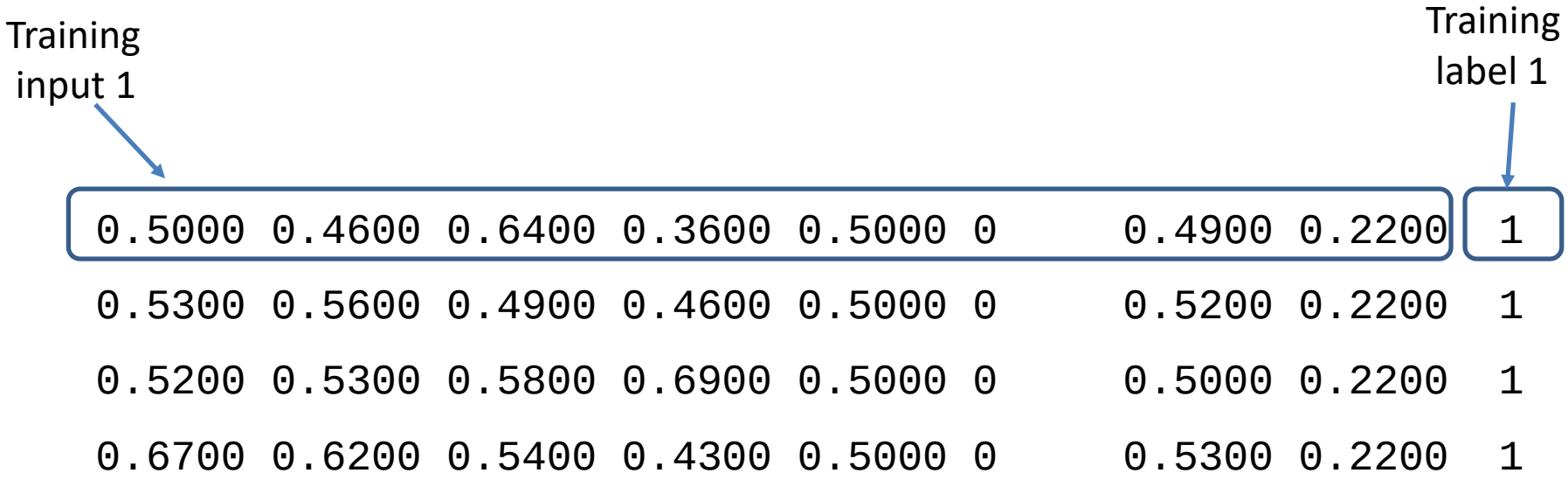
Programming Assignment - Datasets

- The assignment introduces three dataset.
- Each dataset has a training set and a test set.
  - These are the first four lines of the training set of the Yeast dataset:

```
0.5000 0.4600 0.6400 0.3600 0.5000 0     0.4900 0.2200  1
0.5300 0.5600 0.4900 0.4600 0.5000 0     0.5200 0.2200  1
0.5200 0.5300 0.5800 0.6900 0.5000 0     0.5000 0.2200  1
0.6700 0.6200 0.5400 0.4300 0.5000 0     0.5300 0.2200  1
```

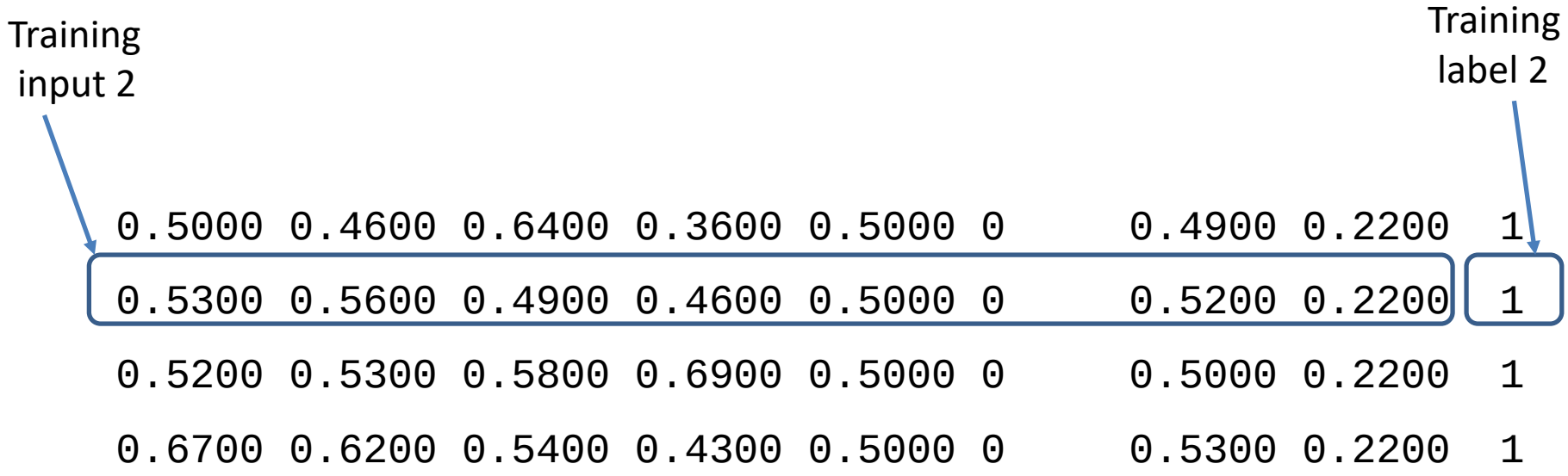
Programming Assignment - Datasets

- Each row is a training example.
- All columns **except for the last column** represent the training input, which is a vector.
- The last column is the class label, which can be a number or a string.



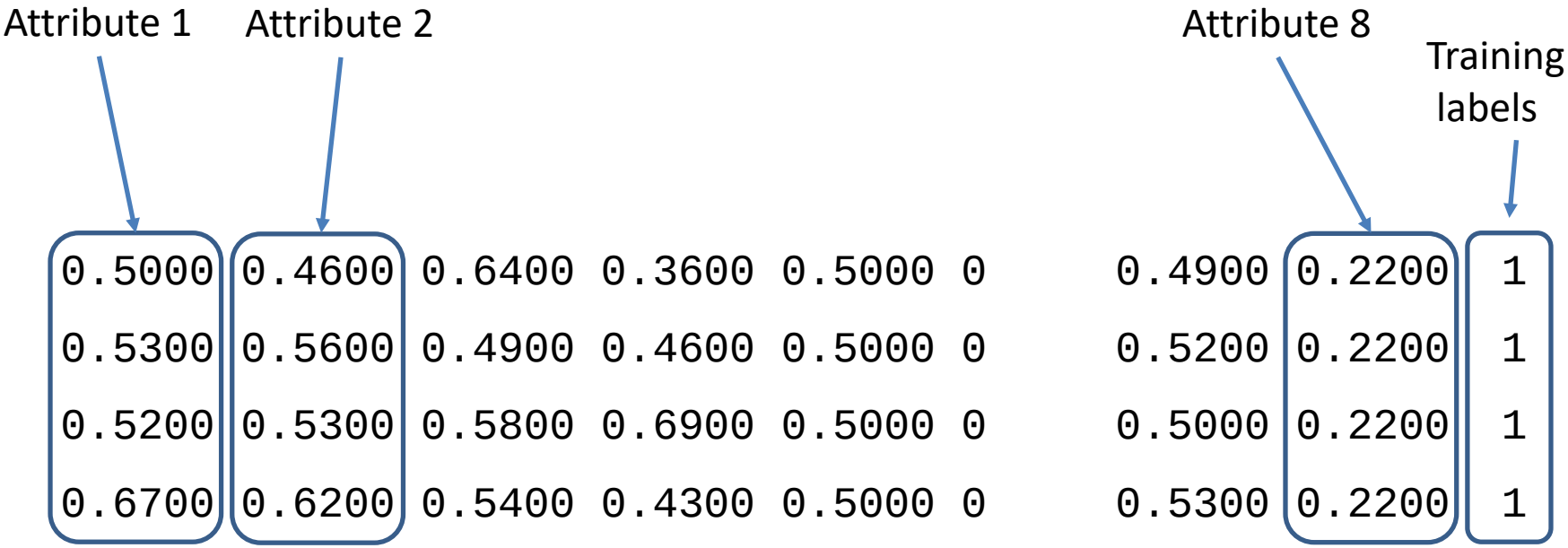
Programming Assignment - Datasets

- Each row is a training example.
- All columns **except for the last column** represent the training input, which is a vector.
- The last column is the class label, which can be a number or a string.



Programming Assignment - Datasets

- Each column (except for the last column) is called a **dimension**, or an **attribute**, or a **feature**.
- In the Yeast dataset, there are 8 attributes/dimensions/features.
- Different datasets have different numbers of attributes.



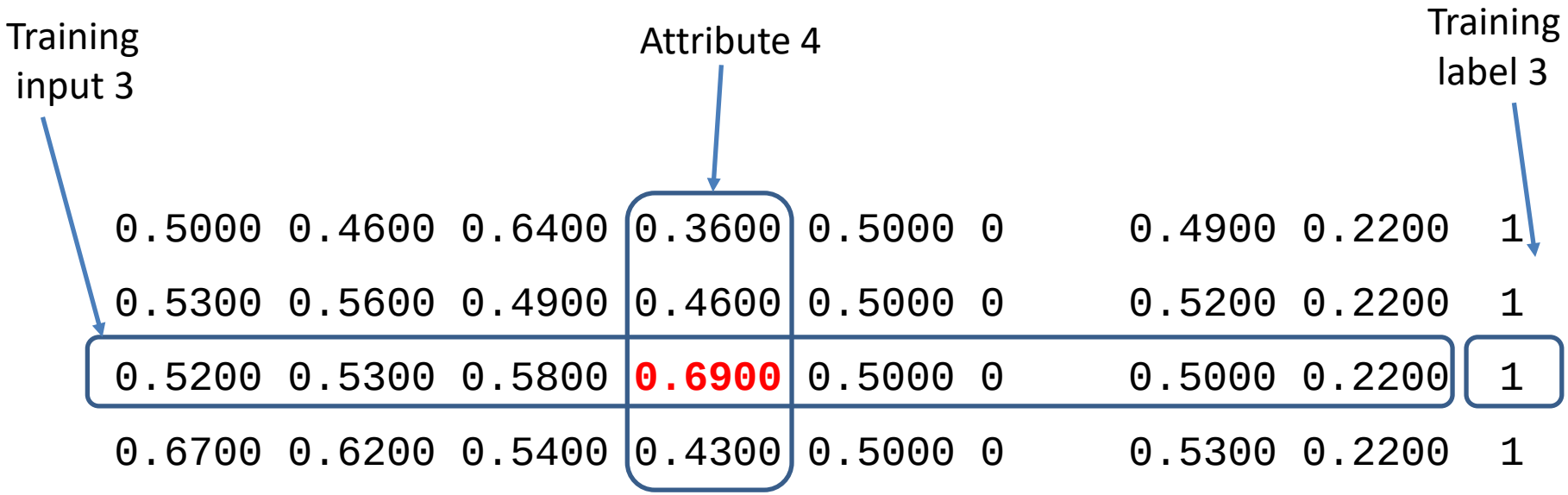
Programming Assignment - Datasets

- So, for example, what is the value for attribute 4 of the third training input?

```
0.5000 0.4600 0.6400 0.3600 0.5000 0     0.4900 0.2200  1
0.5300 0.5600 0.4900 0.4600 0.5000 0     0.5200 0.2200  1
0.5200 0.5300 0.5800 0.6900 0.5000 0     0.5000 0.2200  1
0.6700 0.6200 0.5400 0.4300 0.5000 0     0.5300 0.2200  1
```

Programming Assignment - Datasets

- So, for example, what is the value for attribute 4 of the third training input?
- It is 0.6900.
- We use matrix notation, indices start from 1, not 0.



Class Labels Can Be Strings

- Your code needs to convert class labels to one-hot vectors.
- I suggest a two step approach, as shown on the slides discussing multiclass problems:
  - Step 1: map class labels to consecutive integers that start from 1.
  - Step 2: map integer class labels to one-hot vectors.
  - For easy reference, the next few slides are a copy of the slides we have seen before, that describe these two steps.

Converting to One-Hot-Vectors

- Suppose we have this training set:
  - $\boldsymbol{x}_1 = (0.5, 2.4, 8.3, 1.2, 4.5)^T, \quad q_1 = \text{dog},$
  - $\boldsymbol{x}_2 = (3.4, 0.6, 4.4, 6.2, 1.0)^T, \quad q_2 = \text{dog},$
  - $\boldsymbol{x}_3 = (4.7, 1.9, 6.7, 1.2, 3.9)^T, \quad q_3 = \text{cat},$
  - $\boldsymbol{x}_4 = (2.6, 1.3, 9.4, 0.7, 5.1)^T, \quad q_4 = \text{fox},$
  - $\boldsymbol{x}_5 = (8.5, 4.6, 3.6, 2.0, 6.2)^T, \quad q_5 = \text{cat},$
  - $\boldsymbol{x}_6 = (5.2, 8.1, 7.3, 4.2, 1.6)^T, \quad q_6 = \text{fox},$
- Step 1:
  - Generate new class labels s_n, where classes are numbered sequentially starting from 1. Thus, in our example, the class labels become 1, 2, 3.

Converting to One-Hot-Vectors

- Suppose we have this training set:
  - $\boldsymbol{x}_1 = (0.5, 2.4, 8.3, 1.2, 4.5)^T, \quad q_1 = \text{dog}, \quad s_1 = 1$
  - $\boldsymbol{x}_2 = (3.4, 0.6, 4.4, 6.2, 1.0)^T, \quad q_2 = \text{dog}, \quad s_2 = 1$
  - $\boldsymbol{x}_3 = (4.7, 1.9, 6.7, 1.2, 3.9)^T, \quad q_3 = \text{cat}, \quad s_3 = 2$
  - $\boldsymbol{x}_4 = (2.6, 1.3, 9.4, 0.7, 5.1)^T, \quad q_4 = \text{fox}, \quad s_4 = 3$
  - $\boldsymbol{x}_5 = (8.5, 4.6, 3.6, 2.0, 6.2)^T, \quad q_5 = \text{cat}, \quad s_5 = 2$
  - $\boldsymbol{x}_6 = (5.2, 8.1, 7.3, 4.2, 1.6)^T, \quad q_6 = \text{fox}, \quad s_6 = 3$
- Step 1:
  - Generate new class labels s_n, where classes are numbered sequentially starting from 1. Thus, in our example, the class labels become 1, 2, 3.

Converting to One-Hot-Vectors

- Training set:
  - $\boldsymbol{x}_1 = (0.5, 2.4, 8.3, 1.2, 4.5)^T$, $s_1 = 1$ $\quad\quad$ $\boldsymbol{t}_1 = (?, ?, ?)^T$
  - $\boldsymbol{x}_2 = (3.4, 0.6, 4.4, 6.2, 1.0)^T$, $s_2 = 1$ $\quad\quad$ $\boldsymbol{t}_2 = (?, ?, ?)^T$
  - $\boldsymbol{x}_3 = (4.7, 1.9, 6.7, 1.2, 3.9)^T$, $s_3 = 2$ $\quad\quad$ $\boldsymbol{t}_3 = (?, ?, ?)^T$
  - $\boldsymbol{x}_4 = (2.6, 1.3, 9.4, 0.7, 5.1)^T$, $s_4 = 3$ $\quad\quad$ $\boldsymbol{t}_4 = (?, ?, ?)^T$
  - $\boldsymbol{x}_5 = (8.5, 4.6, 3.6, 2.0, 6.2)^T$, $s_5 = 2$ $\quad\quad$ $\boldsymbol{t}_5 = (?, ?, ?)^T$
  - $\boldsymbol{x}_6 = (5.2, 8.1, 7.3, 4.2, 1.6)^T$, $s_6 = 3$ $\quad\quad$ $\boldsymbol{t}_6 = (?, ?, ?)^T$
- Step 2: Convert each label s_n to a **one-hot vector** $\boldsymbol{t}_n$.
  - Vector $\boldsymbol{t}_n$ has as many dimensions as the number of classes.
    - In our example we have three classes, so each $\boldsymbol{t}_n$ is 3-dimensional.
  - If $s_n = i$, then set the i-th dimension of t_n to 1.
  - Otherwise, set the i-th dimension of t_n to 0.

Converting to One-Hot-Vectors

- Training set:
  - $\boldsymbol{x}_1 = (0.5, 2.4, 8.3, 1.2, 4.5)^T$, $s_1 = 1$ $\quad\quad$ $\boldsymbol{t}_1 = (1, 0, 0)^T$
  - $\boldsymbol{x}_2 = (3.4, 0.6, 4.4, 6.2, 1.0)^T$, $s_2 = 1$ $\quad\quad$ $\boldsymbol{t}_2 = (1, 0, 0)^T$
  - $\boldsymbol{x}_3 = (4.7, 1.9, 6.7, 1.2, 3.9)^T$, $s_3 = 2$ $\quad\quad$ $\boldsymbol{t}_3 = (0, 1, 0)^T$
  - $\boldsymbol{x}_4 = (2.6, 1.3, 9.4, 0.7, 5.1)^T$, $s_4 = 3$ $\quad\quad$ $\boldsymbol{t}_4 = (0, 0, 1)^T$
  - $\boldsymbol{x}_5 = (8.5, 4.6, 3.6, 2.0, 6.2)^T$, $s_5 = 2$ $\quad\quad$ $\boldsymbol{t}_5 = (0, 1, 0)^T$
  - $\boldsymbol{x}_6 = (5.2, 8.1, 7.3, 4.2, 1.6)^T$, $s_6 = 3$ $\quad\quad$ $\boldsymbol{t}_6 = (0, 0, 1)^T$
- Step 2: Convert each label s_n to a **one-hot vector** $\boldsymbol{t}_n$.
  - Vector $\boldsymbol{t}_n$ has as many dimensions as the number of classes.
    - In our example we have three classes, so each $\boldsymbol{t}_n$ is 3-dimensional.
  - If $s_n = i$, then set the i-th dimension of t_n to 1.
  - Otherwise, set the i-th dimension of t_n to 0.