# 生物大数据课程电子病历项目

## 题目

电子病历在医学中的应用越来越广泛，其包含了海量的医学实践信息，被认为是一种可供挖掘的宝库。

《diabetic\_data.csv》文档中给出了约10万糖尿病患者的电子病历的一部分，该数据涵盖了1999-2008年10年来全美130个医院的糖尿病患者住院信息。数据共包含55个特征，文档《Table 1-Feature Description.pdf》中给出了每一个特征的说明，而《IDs\_mapping.csv》中给出了部分特征的编码说明。

本题目要求同学们以前101700例样本作为训练集，以后66例样本作为测试集，每输入1例测试集样本，要求输出训练集中与该测试样本最相似的前5个样本，并按相似度从高到低排序。

在衡量样本的相似性时，并不需要所有特征都加以考虑，本题目要求考虑的特征包括：

|  |  |  |  |
| --- | --- | --- | --- |
| Feature name | Type | Description and values | % missing |
| Gender | Nominal | Values: male, female, and unknown/invalid | 0% |
| Age | Nominal | Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) | 0% |
| Time in hospital | Numeric | Integer number of days between admission and discharge | 0% |
| Glucose serum test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured | 0% |
| A1c test result | Nominal | Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured. | 0% |
| Change of medications | Nominal | Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” | 0% |
| Diabetes medications | Nominal | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” | 0% |
| Readmitted | Nominal | Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission. | 0% |

## Hint

样本之间的相似性可以通过距离来度量，注意连续变量和离散变量适用于不同的距离度量。