# **Chương 2**

## 2.1 Phân tích đề tài

Ngày nay, nhu cầu đi lại, tổ chức sự kiện của con người ngày càng tăng cao. Vì vậy, nếu ta nắm bắt được tình hình thời tiết như lượng mưa, độ ẩm, nhiệt độ,…thì sẽ dễ dàng sắp xếp cũng như tổ chức sự kiện, cuộc vui chơi một cách tốt nhất. Hiểu được nhu cầu cần thiết đó, nhóm em đã quyết định xây dựng bài toán “Dự báo thời tiết” để giúp người dùng có thể dễ dàng có những quyết định tốt nhất cho những dự định sắp tới của mình.

## 2.2 Thuật toán áp dụng

Hiện nay trên thế giới các hệ thống dự báo thời tiết nghiệp vụ được thực hiện bởi nhiều phương pháp khác nhau như: phương pháp thống kê, ứng dụng mạng nơ ron nhân tạo,… .Trong mỗi phương pháp đó lại có những ưu điểm và nhược điểm riêng. Việc ứng dụng phương pháp nào là tùy thuộc vào tài nguyên hệ thống, trình độ của người dự báo, yêu cầu và mục đích của người xây dựng hệ thống,…

Thuật toán KNN (K- nearest neihboor)

### 2.2.1 Định nghĩa

K- nearest neihboor (KNN) là một trong những thuật toán học có giám sát đơn giản nhất trong Machine Learning. Ý tưởng của KNN là tìm ra output của dữ liệu dựa trên thông tin của những dữ liệu training gần đó nhất.

Phương pháp phân lớp *k*-láng giềng gần nhất (*k*-NN) ban đầu dựa trên khoảng cách Euclidean để xác định đâu là các láng giềng gần nhất của mẫu dữ liệu cần phân lớp. Do vậy, dữ liệu thường là kiểu số. Tuy nhiên, việc lựa chọn một độ đo khoảng cách phù hợp có thể mở rộng phương pháp này cho các dữ liệu có kiểu phi số hoặc hỗn hợp.

### 2.2.2 Quy trình làm việc

Cho một bộ dữ liệu huấn luyện *X* gồm *n* mẫu: *X(x1, x2, …, xn)*. Mỗi mẫu dữ liệu *xi (i=1,...,n)* gồm *d* thuộc tính dữ liệu và một thuộc tính lớp.

Một mẫu dữ liệu *y(y1, y2,…,yd) chưa* được xác định lớp. Để phân lớp mẫu *y*, ta tiến hành như sau:

**Bước 1:** Tính khoảng cách từ *y* tới *n* mẫu dữ liệu trong bộ dữ liệu huấn luyện, tức là tính *n* khoảng cách *d(y, xi), i*=1,…,*n*.

**Bước 2:** Lựa chọn ra *k* mẫu dữ liệu trong bộ dữ liệu huấn luyện “*gần*” với *y* nhất, tức có khoảng cách tới *y* là nhỏ nhất.

**Bước 3:** Xác định lớp cho *y*. Lớp của mẫu *y* được xác định là lớp xuất hiện nhiều nhất trong số các giá trị lớp của *k* mẫu gần với *y* nhất đã tìm được ở Bước 2.

Khi sử dụng phương pháp này, ta cần lưu ý một số điểm sau:

Số láng giềng gần nhất (*k*) được coi như đầu vào của thuật toán, tức là cần chọn một giá trị cho *k* trước khi thực hiện thuật toán. Hiển nhiên là số *k* được chọn sẽ có thể ảnh hưởng tới chất lượng của giải thuật. Một câu hỏi đặt ra là: với một bộ dữ liệu cho trước, số *k* được chọn như thế nào để cho kết quả thuật toán tốt nhất? Câu hỏi này mở ra một lĩnh vực nghiên cứu sôi động và các kết quả hầu như đều chưa đưa ra được câu trả lời thỏa đáng. Nói chung, giá trị của *k* phụ thuộc mạnh vào bộ dữ liệu huấn luyện mà ta có. Một số phương pháp phổ biến để xác định số *k* có thể kể tới như *cross-validation* hoặc *bootstrapping*…Chúng thực chất là phương pháp thử với nhiều giá trị *k* khác nhau và lựa chọn ra một số *k* tối ưu đối với một bộ dữ liệu huấn luyện. Trong một số ứng dụng, người ta sử dụng số *k* mặc định bằng 1.

Độ đo khoảng cách cần phải phù hợp với dữ liệu hiện có. Hiển nhiên, độ đo này phải phù hợp với kiểu của dữ liệu. Ngoài ra, với một kiểu dữ liệu cụ thể, lại có thể có nhiều độ đo khoảng cách cho kiểu dữ liệu đó. Lựa chọn độ đo nào trong số các độ đo này là tùy thuộc vào bài toán cụ thể.

### 2.2.3 Ưu điểm, nhược điểm của thuật toán

\* Ưu điểm:

- Dễ sử dụng và cà đặt.

- Việc dự đoán kết quả của dữ liệu mới dễ dàng.

- Độ phức tạp tính toán nhỏ.

\* Nhược điểm

- KNN nhiễu dễ đưa ra kết quả không chính xác khi k nhỏ.

- Cần thời gian lưu training set, khi dữ liệu training và test tăng lên nhiều sẽ mất nhiều thời gian tính toán.

## 3. Bộ dữ liệu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Max Temperature | Min Temperature | Wind Speed | Cloud Cover | Relative Humidity | Conditions |
| 35.9 | 25 | 12.9 | 11.6 | 67.11 | Rain |
| 36 | 27.5 | 12.1 | 11.8 | 68.15 | Rain |
| 35.5 | 28 | 11.2 | 21.7 | 73.05 | Rain |
| 35 | 28 | 19.4 | 27.7 | 71.29 | Rain, Partially cloudy |
| 36 | 28 | 15.7 | 22.1 | 69.52 | Clear |
| 35.7 | 26.5 | 25.9 | 37.2 | 72.42 | Rain, Partially cloudy |
| 28.1 | 25 | 12.1 | 53.2 | 85.11 | Rain, Partially cloudy |
| 31.2 | 25 | 14.8 | 49 | 81.74 | Rain, Partially cloudy |
| 31 | 26 | 19.4 | 67.6 | 80.83 | Rain, Partially cloudy |
| 33.7 | 26.5 | 14.8 | 55.4 | 80.11 | Rain, Partially cloudy |
| 33.2 | 26 | 14.3 | 30.1 | 79.21 | Rain, Partially cloudy |
| 33.5 | 25.5 | 14.8 | 45.1 | 82.51 | Rain, Partially cloudy |
| 32.7 | 25 | 18.4 | 55.3 | 81.14 | Rain, Partially cloudy |
| 33.5 | 26 | 15.8 | 47.6 | 79.8 | Rain, Partially cloudy |
| 33.5 | 26 | 18.4 | 41.4 | 78.08 | Rain, Partially cloudy |
| 34.5 | 27 | 23.2 | 32 | 76.82 | Rain, Partially cloudy |
| 34.5 | 27 | 15.7 | 32.5 | 72.4 | Partially cloudy |
| 29 | 26 | 23.2 | 55.3 | 67.26 | Rain, Partially cloudy |
| 25.5 | 24 | 22.3 | 75.6 | 81.14 | Rain, Overcast |
| 26.7 | 24 | 10.3 | 65.7 | 87.66 | Rain, Partially cloudy |
| 32.5 | 25 | 17.5 | 57 | 80.66 | Partially cloudy |
| 31.8 | 25 | 15.7 | 53.9 | 85.55 | Rain, Partially cloudy |
| 30.9 | 25 | 18.4 | 56.2 | 85.23 | Rain, Partially cloudy |
| 28.5 | 23 | 18.4 | 66.9 | 80.03 | Rain, Partially cloudy |
| 33 | 25 | 16.6 | 53.2 | 75.99 | Rain, Partially cloudy |
| 32.7 | 25 | 16.7 | 56.6 | 79.28 | Rain, Partially cloudy |
| 28.3 | 24 | 21.4 | 67 | 83.47 | Rain, Partially cloudy |
| 30 | 24 | 13 | 65.2 | 81.15 | Rain, Partially cloudy |
| 33 | 25.9 | 16.6 | 44.7 | 77.18 | Rain, Partially cloudy |
| 34.1 | 26 | 16.6 | 37.7 | 75.58 | Rain, Partially cloudy |
| 31 | 25 | 16.6 | 50.9 | 81.67 | Rain, Partially cloudy |
| 32 | 26 | 13.9 | 48 | 79.83 | Rain, Partially cloudy |
| 31 | 25.5 | 13.9 | 48.2 | 87.3 | Rain, Partially cloudy |
| 33 | 25 | 13.9 | 50.8 | 80.45 | Rain, Partially cloudy |
| 32.5 | 26 | 18.5 | 46.2 | 82.36 | Rain, Partially cloudy |
| 29.2 | 24 | 18.5 | 39 | 72.02 | Rain, Partially cloudy |
| 28.1 | 23 | 20.5 | 56.5 | 67.45 | Partially cloudy |
| 27.4 | 22 | 24.1 | 39 | 61.22 | Partially cloudy |
| 27.4 | 21 | 21.4 | 22.7 | 60.35 | Clear |
| 27.1 | 20 | 24.1 | 24.9 | 61.17 | Clear |
| 27.5 | 20.5 | 17.5 | 27.6 | 59.47 | Partially cloudy |
| 29.3 | 21.5 | 19.4 | 19.9 | 62.86 | Clear |
| 30.5 | 25 | 21.4 | 28.8 | 56.39 | Partially cloudy |
| 27 | 22.5 | 23.2 | 93.8 | 83.82 | Rain, Overcast |
| 25 | 22.7 | 19.4 | 93.2 | 85.94 | Rain, Overcast |
| 27 | 22.8 | 25 | 68 | 72.69 | Rain, Partially cloudy |
| 24.5 | 21 | 21.4 | 34 | 60.1 | Rain, Partially cloudy |
| 24 | 20.5 | 23.2 | 52.9 | 69.63 | Rain, Partially cloudy |
| 23 | 21 | 18.4 | 32 | 66.07 | Rain, Partially cloudy |
| 24.5 | 21.4 | 15.7 | 26.3 | 57.91 | Rain, Partially cloudy |
| 27.5 | 21.5 | 21.2 | 18.1 | 54.08 | Clear |
| 29 | 18 | 20.5 | 7.2 | 51.14 | Clear |
| 29 | 17 | 19.4 | 1.7 | 54.61 | Clear |
| 28.5 | 16.5 | 13 | 5.8 | 59.82 | Clear |
| 24.2 | 19 | 14.8 | 71.1 | 71.02 | Rain, Partially cloudy |
| 24 | 22 | 12.1 | 86.9 | 80.24 | Rain, Overcast |
| 29.8 | 22 | 12.1 | 34.2 | 80.55 | Rain, Partially cloudy |
| 30.5 | 22.5 | 20.3 | 46.9 | 74.48 | Rain, Partially cloudy |
| 24 | 22 | 23.2 | 86.3 | 76.04 | Rain, Overcast |
| 22 | 20 | 12.1 | 95.5 | 91.87 | Rain, Overcast |
| 24.5 | 20.7 | 9.4 | 72.6 | 86.18 | Rain, Partially cloudy |
| 28.4 | 21 | 11.2 | 58.2 | 81.98 | Partially cloudy |
| 28 | 22 | 19.4 | 39.8 | 78.98 | Rain, Partially cloudy |
| 24.7 | 21 | 15.7 | 36.4 | 59.71 | Rain, Partially cloudy |
| 24.8 | 20.5 | 15.7 | 31 | 59.81 | Partially cloudy |
| 26.7 | 19 | 10.3 | 33.7 | 70.67 | Partially cloudy |
| 26 | 22 | 10.1 | 25.4 | 74.95 | Partially cloudy |
| 28.2 | 22.1 | 20.3 | 20.8 | 61.08 | Clear |
| 29 | 21.5 | 17.6 | 5.6 | 53.88 | Clear |
| 28 | 20.3 | 19.4 | 0.1 | 49.5 | Clear |
| 27.2 | 17 | 17.5 | 13 | 48.97 | Clear |
| 27 | 18.5 | 12.1 | 23 | 51.92 | Clear |
| 28.3 | 19 | 13.9 | 5.2 | 52.17 | Clear |
| 29 | 18.5 | 16.6 | 1 | 51.91 | Clear |
| 28 | 15.5 | 18.5 | 11.8 | 52.47 | Clear |
| 19.5 | 18.7 | 18.5 | 96.2 | 90.15 | Rain, Overcast |
| 21 | 19 | 8.5 | 93.8 | 94.97 | Rain, Overcast |
| 27.1 | 20.7 | 18.5 | 77.7 | 89.35 | Rain, Overcast |
| 28.7 | 22.8 | 13.9 | 66.5 | 81.48 | Partially cloudy |
| 28 | 22.8 | 12.1 | 68.7 | 83.51 | Rain, Partially cloudy |
| 29 | 23 | 16.6 | 72.4 | 81.42 | Partially cloudy |
| 30 | 24 | 19.4 | 66.8 | 81.35 | Partially cloudy |
| 29 | 23 | 14.8 | 63.2 | 81.67 | Partially cloudy |
| 29 | 23.5 | 16.6 | 52.8 | 76.66 | Partially cloudy |
| 26.5 | 23 | 15.7 | 68.8 | 75.58 | Partially cloudy |
| 27 | 22 | 13.9 | 53.4 | 76.99 | Partially cloudy |
| 28.5 | 22 | 12.1 | 53.4 | 73.49 | Partially cloudy |
| 27 | 19 | 19.4 | 35.3 | 71.01 | Partially cloudy |
| 22.5 | 18 | 20.5 | 64.5 | 65.73 | Partially cloudy |
| 22.1 | 17 | 21.4 | 22 | 52.63 | Clear |
| 23.2 | 16 | 22.3 | 14.6 | 48.87 | Clear |
| 21 | 17.5 | 18.4 | 36.8 | 52.27 | Partially cloudy |
| 21 | 16 | 14.8 | 62.4 | 64.18 | Partially cloudy |
| 22 | 18 | 18.4 | 33.9 | 55.18 | Partially cloudy |
| 20.3 | 14.5 | 18.5 | 43.8 | 54.88 | Partially cloudy |
| 20 | 13 | 20.5 | 21.3 | 57.21 | Clear |
| 22 | 11 | 10.3 | 0.1 | 64.98 | Clear |
| 23 | 15.1 | 15.7 | 41.6 | 63.4 | Partially cloudy |