ITF: **Information Technology Faculty**
Thuyloi University

K lân cận gần nhất

Trình bày: PGS.TS Nguyễn Hữu Quỳnh

Bài giảng được dựa trên giáo trình machine learning cơ bản

Giới thiệu

Có một anh bạn chuẩn bị đến ngày thi cuối kỳ (được mở tài liệu):

- Anh ta không chịu ôn tập để hiểu ý nghĩa của từng bài học và mối liên hệ giữa các bài.
- Anh thu thập tất cả các tài liệu trên lớp, bao gồm:
  - ghi chép bài giảng (lecture notes),
  - các slides và
  - bài tập về nhà + lời giải.
- Để cho chắc, anh ta ra thư viện và các quán Photocopy quanh trường mua hết tất cả các loại tài liệu liên quan.
- Cuối cùng, anh ta thu thập được một chồng cao tài liệu để mang vào phòng thi.

Giới thiệu

- Vào ngày thi, anh tự tin mang chồng tài liệu vào phòng thi. Aha, đề này ít nhất mình phải được 8 điểm:
  - Câu 1 giống hệt bài giảng trên lớp.
  - Câu 2 giống hệt đề thi năm ngoái mà lời giải có trong tập tài liệu mua ở quán Photocopy.
  - Câu 3 gần giống với bài tập về nhà.
  - Câu 4 trắc nghiệm thậm chí cậu nhớ chính xác ba tài liệu có ghi đáp án.
  - Câu cuối cùng, 1 câu khó nhưng anh đã từng nhìn thấy, chỉ là không nhớ ở đâu thôi.

Giới thiệu

- Kết quả cuối cùng, cậu ta được 4 điểm, vừa đủ điểm qua môn:
  - Cậu làm chính xác câu 1 vì tìm được ngay trong tập ghi chú bài giảng.
  - Câu 2 cũng tìm được đáp án nhưng lời giải của quán Photocopy sai!
  - Câu ba thấy gần giống bài về nhà, chỉ khác mỗi một số thôi, cậu cho kết quả giống như thế luôn, vậy mà không được điểm nào.
  - Câu 4 thì tìm được cả 3 tài liệu nhưng có hai trong đó cho đáp án A, cái còn lại cho B. Cậu chọn A và được điểm.
  - Câu 5 thì không làm được dù còn tới 20 phút, vì tìm mãi chẳng thấy đáp án đâu - nhiều tài liệu quá cũng mệt!!
- Đó là học lười (lazy learning)

K-nearest neighbor

- Khi training, thuật toán này *không học* một điều gì từ dữ liệu training
- Mọi tính toán được thực hiện khi nó cần dự đoán kết quả của dữ liệu mới.
- K-nearest neighbor có thể áp dụng được vào cả hai loại của bài toán Supervised learning là Classification và Regression.

K-nearest neighbor

- Một vài khái niệm tương ứng người-máy:

Ngôn ngữ người	Ngôn ngữ Máy Học	in Machine Learning
Câu hỏi	Điểm dữ liệu	Data point
Đáp án	Đầu ra, nhãn	Output, Label
Ôn thi	Huấn luyện	Training
Tập tài liệu mang vào phòng thi	Tập dữ liệu tập huấn	Training set
Đề thi	Tập dữ liểu kiểm thử	Test set
Câu hỏi trong dề thi	Dữ liệu kiểm thử	Test data point
Câu hỏi có đáp án sai	Nhiễu	Noise, Outlier
Câu hỏi gần giống	Điểm dữ liệu gần nhất	Nearest Neighbor

KNN trong Classification

- label của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set.
- Label của một test data:
  - có thể được quyết định bằng major voting giữa các điểm gần nhất
  - hoặc nó có thể được suy ra bằng cách đánh trọng số khác nhau cho mỗi trong các điểm gần nhất đó rồi suy ra label

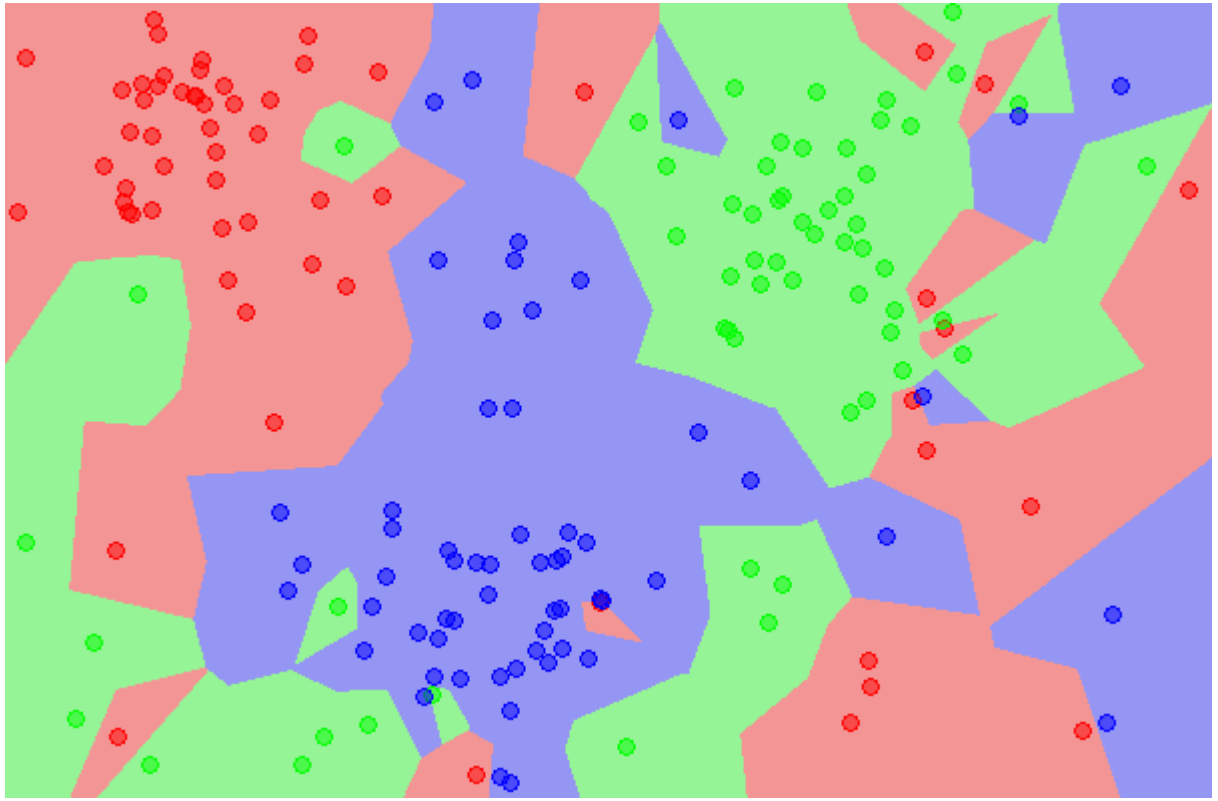
KNN trong Regresssion

- đầu ra của một điểm dữ liệu:
- sẽ bằng chính đầu ra của điểm dữ liệu đã biết gần nhất (trong trường hợp K=1)
- hoặc là trung bình có trọng số của đầu ra của những điểm gần nhất,
- hoặc bằng một mối quan hệ dựa trên khoảng cách tới các điểm gần nhất đó

Tóm tắt về KNN

- KNN là thuật toán đi tìm đầu ra của một điểm dữ liệu mới bằng cách *chỉ* dựa trên thông tin của K điểm dữ liệu trong training set gần nó nhất (K-lân cận),
- KNN *không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu*.
- KNN phải *nhớ* tất cả các điểm dữ liệu training: việc này không được lợi về cả bộ nhớ và thời gian tính toán

Ví dụ về KNN trong classification với K = 1.



Khoảng cách trong không gian vector

- Trong không gian một chiều:, khoảng cách giữa hai điểm là trị tuyệt đối giữa hiệu giá trị của hai điểm đó.
- Trong không gian nhiều chiều:
  - khoảng cách giữa hai điểm có thể được định nghĩa bằng nhiều hàm số khác nhau,
  - độ dài đường thẳng nổi hai điểm chỉ là một trường hợp đặc biệt trong đó.