MSSP 608 - Recitation 3

Chetan Parthiban (chetanp@sas.upenn.edu)

We've gotten a lot of interest in GitHub -- but we'll cover it after the final lectures

Note from the last homeworks:

Don’t Overfit to the Test Set!!! Use a validation set.

Today's Topics

- Naive Bayes, SVMs, KNNs
- Word Embeddings/Domain Transfer
- A couple tricks in colab

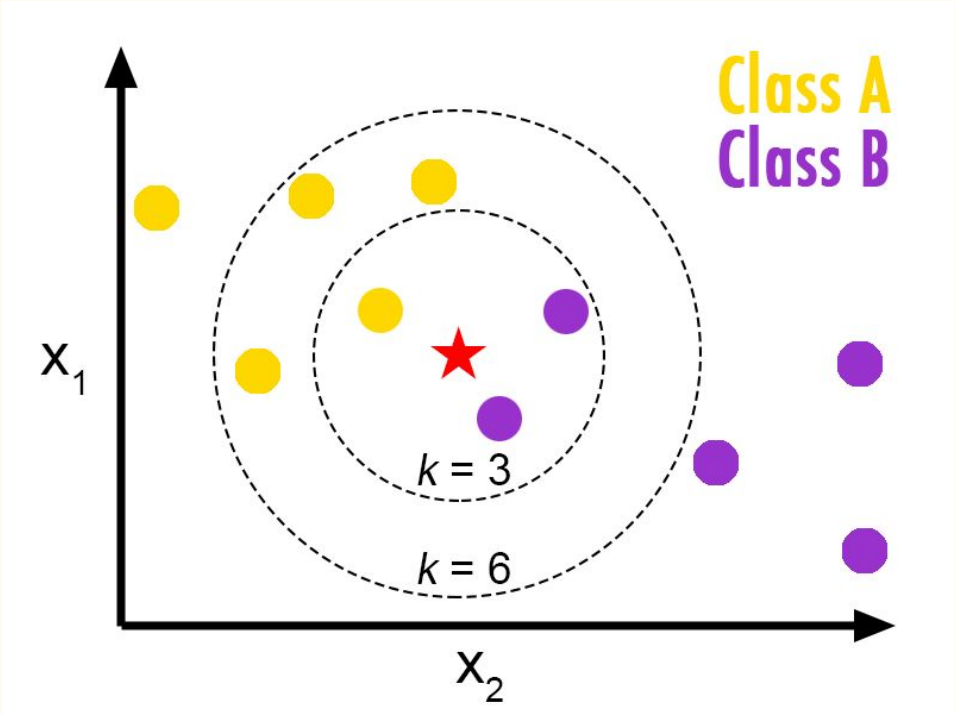
Naive Bayes, KNNs, SVMs

Naive Bayes - On the Board

k-Nearest Neighbors

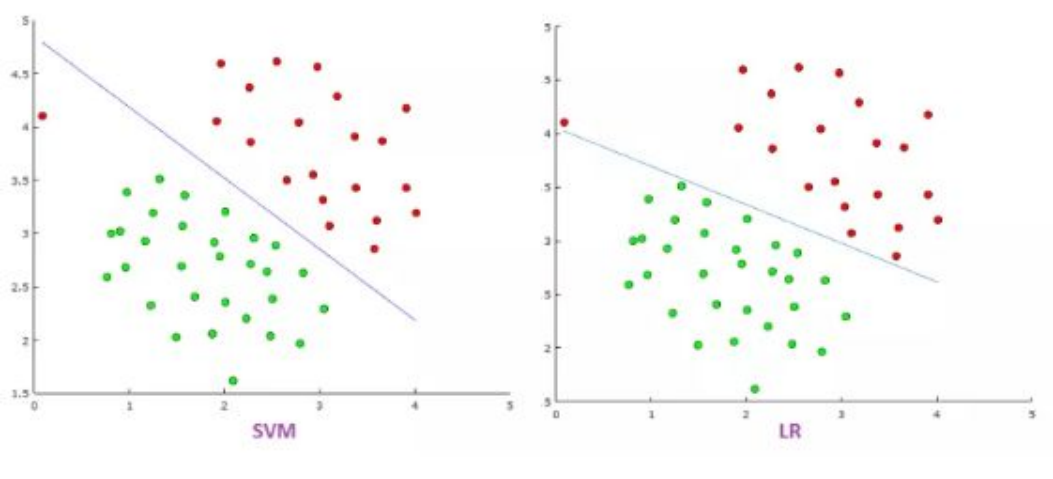
Probably the most intuitive classifier -- just pick the class based on similar points

In practice, this can be expensive to implement since it takes longer to run the more data you have



Support Vector Machines

Similar to logistic regression -- it draws a plane through the data



Differences

1) It tries to maximize the distance between the point and the plane, not just classify correctly
2) No output “probabilities” just the class

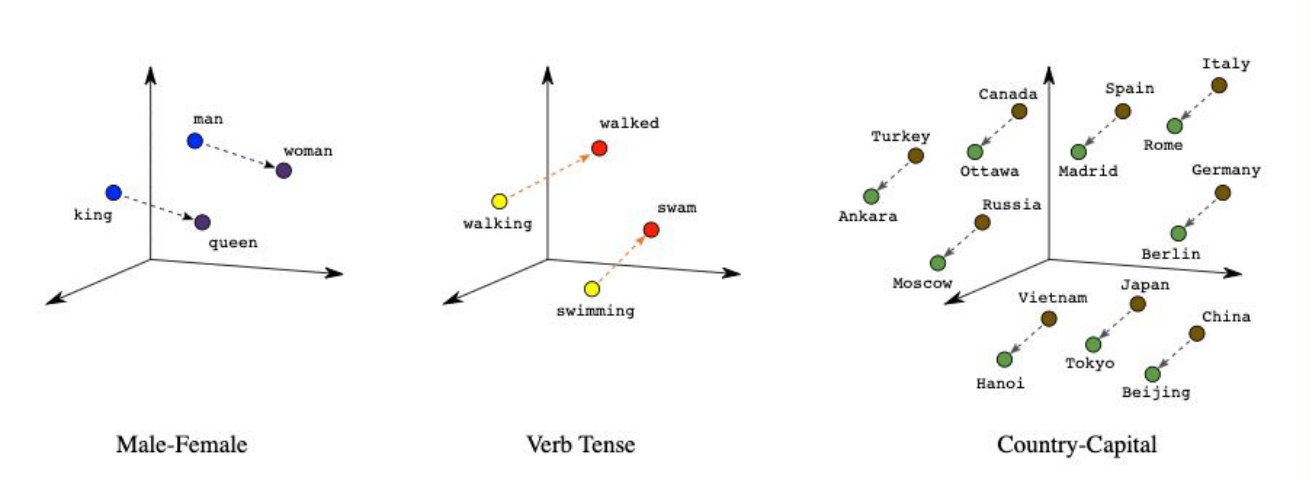
Word Embeddings & Domain Transfer

What are embeddings?

We learn a representation for the “meaning” of the words through their contexts

Create a map between the words and 300 dimensional vector space

None of the dimensions have intrinsic meaning, but certain directions can capture semantic ideas (e.g. gender)



Transfer Learning & Domain Transfer

In the real world we commonly encounter the following scenario.

1) We have lots of Data A, which is easier to obtain
2) We have a small amount of Data B, which is hard to obtain
3) Data A and B are somewhat related
4) Our task of interest needs Data B, but could leverage some information from Data A

Domain Transfer is the way that we solve this problem

For example, learning to understand the language is a key step in any NLP system that can be shared regardless of the end goal task

Power of Embeddings + Transfer Learning -- AI Dungeon 2

Tricks in Google Colab

A few tricks to save you a bit of time (demo)

- Table of Contents
- Code snippets and Interactive Components
- File System + File Previews
- Code auto-complete and documentation
- Multi-cursor