DATA 5610

- Alpaydin Chp 5: Multivariate Methods
  - 5.1 Multivariate Data
  - 5.2 Parameter Estimation
  - 5.3 Estimation of Missing Values
  - 5.4 Multivariate Normal Distribution
  - 5.5 Multivariate classification
  - 5.6 Tuning complexity
  - 5.7 Discrete Features

- ISLR2
  - Chp 4: Classification

- [redacted] Wilmott Chp 5 Naive Bayes Classifier

Naive Bayes Classifier → models the joint pdf $p(x, y)$

- Good intro to "generative" classification

- Setup:

  - Given Data $= [(x^{(1)}, y), \ldots, (x^{(n)}, y)]$

    $x^{(i)} = (x^{(i)}_1, \ldots, x^{(i)}_d)$ so $x^{(i)} \in \mathbb{R}^d$, $y_i \in Y$

    where $Y = \{1, \ldots, m\}$

  - Assume a family of distributions P_θ s.t.

    $$P_\theta(x, y) = P_\theta(x|y)\, p(y)$$

    $$= P_\theta(x_1|y) \cdots p(x_d|y)\, P_\theta(y)$$

    * This key assumption is why it is called "Naive"

- Let $(x^{(1)}, y_1), \ldots, (x^{(n)}, y_n) \sim P_\theta$ iid for some θ

(If $(X, y) \sim P_\theta$, then $X_1, \ldots, X_d$ are independent given Y)

("Conditional Independ. Assumption") → "Independent Feature Model"

- $X_1, \ldots, X_d$ the "features"

- Goal: For some new $X \in \mathbb{R}^d$, predict its Y (classify)

- Algorithm:
  - Estimate θ (MLE, MAP, etc.) from Data
  - Compute $\hat{y} = \underset{y \in Y}{\operatorname{argmax}}\ P_{\hat{\theta}}(y \mid x)$

$\hat{\theta}$: estimated

$$\hat{y} = \underset{(y)}{\operatorname{argmax}} \left[\frac{P_{\hat{\theta}}(x|y)\, P_{\hat{\theta}}(y)}{P_{\hat{\theta}}(x)} \right]$$ using Bayes' Rule

- Does not depend on denom., so

$$\hat{y} = \underset{y}{\operatorname{argmax}} \left[P_{\hat{\theta}}(x|y)\, P_{\hat{\theta}}(y) \right]$$ will select the same $\hat{y}$

$$= \underset{y}{\operatorname{argmax}}\; P_{\hat{\theta}}(x_1|y) \cdots P_{\hat{\theta}}(x_d|y)\, P_{\hat{\theta}}(y)$$

- The "Bayes" part of NBC is b/c it is a Bayes' Estimator

What is it used For?

- Peter Norvig: Often start with NBC and don't need to improve on it
- Classification, text, sentiment
- NLP: Natural Language Processing (getting the computer to "understand" humans)
- Email spam detection (your email prolly uses NBC)
- "Good" or "bad" news
- Predict which direction tweets on Twitter will influence election or referendum
- Determine if the tweets come from a Russian bot

Political Speeches (Wilmott example)

- using Bayes Rule

$$P(\text{Politician is left wing} \mid \text{used the word "comrade"}) =$$

$$\frac{P(\text{used the word "comrade"} \mid \text{left wing})\, P(\text{left wing})}{P(\text{uses the word "comrade"})}$$

- So if a politician uses the word "comrade" in a speech, we can calculate the prob. of him/her being left-wing

- Apply to whole phrases and entire speeches not just single words

- Suppose a speech contains the phrase "Property is theft, Comrade" and we want to know if they are left or right wing

$$P(\text{Left} \mid \text{"Property is theft, Comrade"})$$

and similar for

$$P(\text{Right} \mid \text{"Property is theft, Comrade"})$$

- Now comes the "Naive" part

$$P(\text{"Property is theft, comrade"} \mid \text{Left}) =$$

$$P(\text{"property"} \mid \text{Left}) * P(\text{"theft"} \mid \text{Left}) * P(\text{"Comrade"} \mid \text{Left})$$

- Drop the "is" because it's a stop word

In Symbols

- The data, text we want to classify is called X as a vector consisting of x_d for $1 \leq d \leq D$ so that there are D words in the text

- We want to find

$$p(C_k | x)$$

for each of the K classes (political persuasions) C_k

- use MLE or MAP to estimate

Bayes Rule tells us that

$$P(C_k|x) = \frac{P(x|C_k)\,P(C_k)}{P(x)}$$

- when the features are independent this simplifies to

$$P(C_k) \prod_{d=1}^{D} P(x_d|C_k)$$

- This is what we compare for each class
  - We will get $P(x_d|C_k)$ from the data set

- Finally, because multiplying a bunch of possibly (very) small numbers can cause round-off error we take the log transform

$$\ln\left[P(C_k)\right] + \sum_{d=1}^{D} \ln\left[P(x_d \mid C_k)\right]$$

- $\ln\left[P(C_k)\right]$ is the prior
  - We can ignore it (Frequentist)
  - Estimate it from data (Empirical Bayes)
  - Or model it subjectively (Bayesian)

Wilmott's Example

- Churchill "Beaches" speech
- JFK inaugural address
- Benn: Maiden speech as MP
- Thatcher: "The Lady's not for turning" speech
- May: Syria speech
- Corbyn: Post-Brexit-Referendum speech
- Trump: state of the Union speech

* Predict MLK's "I have a dream" speech
* What is your prediction?