CAS CS 365
Lec 7

Probability

1. Real-world problem: how can we classify an email as spam?
    a. Consider the email as a collection of words $w_1, w_2, \ldots, w_n$
        i. Certain words that appear often in all emails (example - "the")
    b. Formulate mathematically our problem:
        i. We are interested in posterior probability Pr(spam|$w_1, w_2, \ldots, w_n$)
    c. Classes = {spam, not spam}
        i. Suppose an email is equally like to be spam or non-spam
    d. We apply Bayes' rule
    e. $$\Pr(\text{spam} \mid w_1, \ldots, w_n) = \frac{\Pr(w_1, \ldots, w_n \mid \text{spam}) \Pr(\text{spam})}{\Pr(w_1, \ldots, w_n \mid \text{spam}) \Pr(\text{spam}) + \Pr(w_1, \ldots, w_n \mid \text{not spam}) \Pr(\text{not spam})}$$
    f. Bad idea → since the probability that all n words appear in spam emails is very unlikely → close to 0, which means that the bayes' probability will be 0 (meaningless information)
2. Bayes Classifier
    a. More generally suppose we have k classes, $\{c_1,...,c_k\}$ with a given prior Pr(C=c_j) = p_j
        i. In our example, k = 2
    b. We have some data D
        i. In our example, the set of words in the e-mail
    c. Suppose we somehow know exactly Pr(C=c_i | D) for each class c_i
    d. Ex)
        i. Pr(lion | photo) = 0.8
        ii. Pr(cat | photo) = 0.15
        iii. Pr(mouse | photo) = 0.05
        iv. Then, given the photo, we are likely to say that the photo shows an image of a lion
    e. What class would you assign to D?
    f. $$c^\star = h_{\text{Bayes}}(\mathcal{D}) = \arg\max_i \Pr(C = i \mid \mathcal{D})$$
    g. Choose the class with the highest percentage
3. Naive Bayes Classifier
    a. A popular classifier is known as Naive Bayes and makes the following conditional independence assumption: (it is an assumption, which is incorrect in real world)

b. In real world, $Pr(w_1, \ldots, w_n \mid c) = Pr(w_1 \mid c) * Pr(w_2 \mid w_1, w_2, c) * \ldots * Pr(w_n \mid w_1 * \ldots * w_{n-1} \mid c)$

c. $$\boxed{\begin{aligned} \Pr(w_1, \ldots, w_n \mid \text{spam}) &= \Pr(w_1 \mid \text{spam}) \cdot \ldots \cdot \Pr(w_n \mid \text{spam}) \\ \Pr(w_1, \ldots, w_n \mid \text{not spam}) &= \Pr(w_1 \mid \text{not spam}) \cdot \ldots \cdot \Pr(w_n \mid \text{not spam}) \end{aligned}}$$

namely words (attributes/features) are conditionally independent given the class

d. Decision rule: output the class c that maximizes the posterior probability

e. $$c^{\star}_{\text{Naive-Bayes}} = \arg\max_{c} \ \Pr(c) \prod_{i=1}^{n} \Pr(w_i \mid c)$$

f. In other words, $Pr(w_1, \ldots, w_n \mid c) = \prod_{i=1}^{n} \Pr(w_i \mid c)$

g. In the current example, we get

$$Pr(\text{spam} \mid \text{dataset}) = \frac{Pr(\text{dataset} \mid \text{spam}) * Pr(\text{spam})}{Pr(\text{dataset})}$$

$$Pr(\text{not spam} \mid \text{dataset}) = \frac{Pr(\text{dataset} \mid \text{not spam}) * Pr(\text{not spam})}{Pr(\text{dataset})}$$

Since Pr(dataset) is equal and exists in both probabilities, we can disregard them (they cancel out)

h. Suppose we have access to a training set, namely a set of emails that are associated with a label {spam, non-spam}

  i. How can we use this information to classify an unseen email?

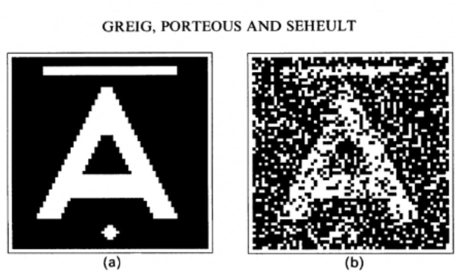
i. We learn the probabilities of Naive Bayes classifier from the training data

j. Prior: $\Pr(\text{spam}) = \dfrac{\#\text{spam emails}}{\#\text{emails}}, \ \Pr(\text{not spam}) = \dfrac{\#\text{not spam emails}}{\#\text{emails}} = 1 - \Pr(\text{spam})$

Likelihood: $\boxed{\Pr(Word = w \mid C = c) = \dfrac{\#(Word = w, Class = c)}{\#\text{emails of Class c}}}$

4. Exact Map Estimation for Binary images

  a. Problem: Given (b) can we infer (a)? In other words, can we restore the image from its corrupted-by-noise version?

  b. 

c. Think of it in real life: matrix with binary numbers (0,1) → 0 if black and 1 if white with some noise
d. Let's be Bayesian
    i. X = (x_1, …, x_n) the original image (shown in figure a) with n = k^2 where k is the number of rows//columns (assuming that the image is square)
    ii. Y = (y_1, …, y_n) the observed corrupted image (shown in figure b)
    iii. This means that there are total of 2^(k^2) matches in total
    iv. We want to find out
    Pr(X | Y) where X and Y are vectors/matrix
    v. Assumption: the records y_1, …, y_n are conditionally independent given x, and each has known conditional density f(y_i | x_i) that depends only on x_i
e. By Bayes' theorem:

$$p(x|y) \propto \underbrace{p(y|x)}_{\text{likelihood:how do we compute it?}} \times \underbrace{p(x)}_{\text{prior:what is a good prior?}}$$

- For example, if given y of {1,1,1,1}, is the value of x = {0,0,0,0} more likely or x = {0,0,1,1} more likely?
- In this question, there is a possibility that that pixel changes from 0 to 1 (black to white) and 1 to 0 (white to black)
- Let's assume they are equal P_{wb} = P_{bw}
- If we go back, Pr({1,1,1,1} | {0,0,1,1}) = (1-P_{wb})^2 * (P_{wb})^2 → two pixels changed and two pixels remained the same

f. Goal: output:

$$\boxed{x^* = \arg\max p(x|y)}$$

g. Likelihood and prior:
    i. Given our assumption, the likelihood function is

$$p(y|x) = \prod_{i=1}^{n} f(y_i|1)^{x_i} f(y_i|0)^{1-x_i}.$$