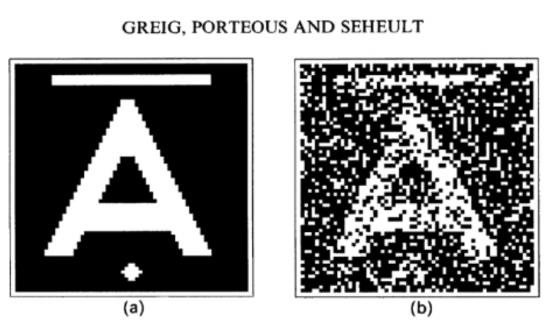
CAS CS 365
Lec 8

Foundations of Data Science

1. Bayesian
   a. $$\mathbf{Pr}[H|D] = \frac{\mathbf{Pr}[D|H]\mathbf{Pr}[H]}{\mathbf{Pr}[D]}, \text{ and } \mathbf{Pr}[D] > 0, \text{ or } \ldots$$
   b. $$\text{posterior} \propto \text{likelihood} \times \text{prior.}$$
   c. Likelihood → Pr[D|H]
   d. Assume a coin gives [HHHT]
      i. Pr[Heads] = ¾
      ii. What is the probability of [HHHT] given probability of heads is p?
         1. Pr[HHHT] = p^3 * (1-p)
         2. Find the maximum p → derivative of p^3(1-p) → 3p^2 - 4p^3 = p^2(3-4p)
         3. To find the maximum, p^2(3-4p) = 0 → p = ¾
   e. Why not maximize likelihood and why posterior (includes prior)?
      i. When you maximize the likelihood, it is trivial (it will output the input)
      ii. When considering posterior, you include prior variable, which is important
2. Problem: Given b, can we infer a? In other words, can we restore the image from its corrupted-by-noise version?
   a. 

      i. $x = (x_1,\ldots,x_n)$ the original image → output
      ii. $y = (y_1,\ldots,y_n)$ the observed corrupted image → input
      iii. Assumption: the records $y_1,\ldots,y_n$ are conditionally independent given x, and each has known conditional density $f(y_i|x_i)$ that depends only on x_i.
      iv. Pr(the event that pixel flips) = p
         1. Example)
         2. Pr([1,1,1,0,0,0,0,0,0] | [1,1,0,1,0,0,0,0,0]) = p^2*(1-p)^7

        3. There are total of 2^9 possible xs given y
        4. If we let p (the probability that pixel changes) = 0.1, likelihood is trivial in this case, because it is easier to to have more (1-p)= 0.9, which is the proportion of not flipped pixels → prior plays an important role
    v. By Bayes' theorem,

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

    vi. Three questions that arise
        1. How do you compute likelihood
        2. How do you calculate prior
        3. How to optimize the x value
    vii. Goal: output $\boxed{x^* = \arg\max p(x|y)}$
        1. To compute the maximum x, we need to compute all possible xs given y → not necessary in this case
b. Likelihood and prior
    i. Likelihood function:

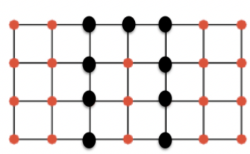
$$p(y|x) = \prod_{i=1}^{n} f(y_i|1)^{x_i} f(y_i|0)^{1-x_i}.$$

If $y_i = x_i$, then 1-p

If y_i != x_i, then p

    ii. Prior function:

$$p(x) \propto \exp\left\{\frac{1}{2}\sum_{i\neq j}\beta_{ij}\left(x_i x_j + (1-x_i)(1-x_j)\right)\right\}$$



i and j are neighbors

The expression is expressing homogeneousness.

x_i	x_j	Value
0	0	1
0	1	0
1	0	0
1	1	1

1. Prior gives a higher probability to the image that is very homogeneous (the image given in A has a pattern that is very homogeneous → the pixel next to a pixel has the same color (two neighboring pixel are likely to be the same color (not discontinuous))
2. Pr(x) is proportional to exp {∑ ($x_i = x_j$)} → if x_i is equal x_j, it means that the neighboring pixel has the same color
3. For an edge (u,v) where val(u) = val(v)

$$x_u x_v + (1 - x_v)(1 - x_u) = x_u^2 + (1 - x_u)^2 = 1.$$

4. On the contrary for an edge where val(u) does not equal val(v)

$$x_u(1 - x_u) + (1 - x_u)x_u = 0.$$

5. Likelihood: trying to fit the data
6. Prior: find homogeneous (neighbor pixels)

iii. Remark: Since are considering the max posterior Pr(x|y) (or equivalently log p(x|y))
   1. Log (p(x|y)) is proportional to log(p(x)) + log p(y|x)
   2. log(p(x)) is the term that is pushing x to be homogeneous
   3. log p(y|x) is fit data y