CSCC11 Week 8 Notes

Review of Class Conditionals:

- Want to model $P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$
- If $P(C_1|x) > P(C_2|x)$, then we classify the input as belonging to class 1.
  If $P(C_1|x) < P(C_2|x)$, then we classify the input as belonging to class 2.
  $P(C_1|x) = P(C_2|x)$ is the decision boundary.
- Can also do $\frac{P(C_1|x)}{P(C_2|x)} > 1$

  or equivalently $\ln\left(\frac{P(C_1|x)}{P(C_2|x)}\right) > 0$
- Recall that $P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}$

  Hence, $\ln\left(\frac{P(C_1|x)}{P(C_2|x)}\right) = \ln\left(\frac{P(x|C_1) \cdot P(C_1)}{P(x|C_2) \cdot P(C_2)}\right)$

  Note: The $P(x)$ term cancels out.
- We can model $P(x|C_1)$, $P(x|C_2)$ as Gaussians. Assuming d features for each input x (I.e. $x \in \mathbb{R}^d$), we have $O(d^2)$ parameters. This is from the Covariance Matrix.

Naive Bayes (NB)

- Naive bayes aims to simplify the estimation problem by assuming that the diff input features (the diff elements of the input vector) are conditionally independent.

I.e. $P(X|C) = \prod_{i=1}^{d} P(X_i|C)$

- With this assumption, rather than estimating 1 d-dimension density, we estimate d 1-dimension densities. This is important bc each 1D Gaussian only has 2 parameters (mean and variance) both of which are scalars. Hence, the model has 2d unknowns. In the Gaussian case, the NB model replaces the dxd covariance matrix by a diagonal matrix. The i^{th} entry is the variance of $X_i|C$.

Discrete Input Features

- In discrete NB, the inputs are a discrete set of features.

- Right now, we'll assume that each input either has or does not have each feature.

- Each data vector is described by a list of discrete features (I.e. $F_{1:d} = [F_1, ..., F_d]$) and for simplicity, we'll assume that each feature is binary (I.e. $f_i = \{0,1\}$).

- Consider this: We want to solve $P(F_1, F_2, F_3 | C=1)$. Without using naive bayes, we would get

Note: This formula comes from the chain rule.

$$P(F_1, F_2, F_3 | C=1) = P(F_1 | F_2, F_3, C=1) \cdot P(F_2 | F_3, C=1) \cdot P(F_3 | C=1)$$

For $P(F_3 | C=1)$, since we know $F_3 = \{0,1\}$, we can model it with 1 number.

For $P(F_2 | F_3, C=1)$, F_2 depends on F_3 and we know that F_3 has 2 possible values, we need to model 2 diff distributions.

For $P(F_1 | F_2, F_3, C=1)$, we need to model 4 diff distributions.

For d-dimensional binary inputs, there are $d(2^d - 1)$ parameters one needs to learn.

With Naive Bayes, only d parameters have to be learned.

This is because $P(F_{1:d} | C=j) = \prod_i P(F_i | C=j)$

- Continuing with NB's way:
  - Let $a_{ij} \equiv P(F_i = 1 | C=j)$
  - Let $b_j \equiv P(C=j)$ ← Prior
  - $P(C=j | F_{1:d}) = \dfrac{P(F_{1:d} | C=j)\, P(C=j)}{P(F_{1:d})}$

$$= \frac{\left(\prod_i P(F_i | C=j)\right)\left(P(C=j)\right)}{\sum_{\ell=1}^{k} P(F_{1:d}, C=\ell)}$$

$$= \frac{\left(\prod_{i:F_i=1} a_{ij} \prod_{i:F_i=0} (1-a_{ij})\right) b_j}{\sum_{\ell=1}^{k} \left(\prod_{i:F_i=1} a_{i\ell} \prod_{i:F_i=0} (1-a_{i\ell})\right) b_\ell}$$

- If we wish to find the class with max posterior prob, we only need to compute the numerator.

- The computation shown on the prev page can lead to underflow.
  To avoid these issues, it's safer to perform the computations in the log-domain:

$$\alpha_j = \left(\sum_{i:F_i=1} \ln a_{ij} + \sum_{i:F_i=0} \ln(1-a_{ij}) \right) + \ln b_j$$

$$\gamma = \min_j \alpha_j$$

$$P(C=j \mid F_{1:d}) = \frac{\exp(\alpha_j - \gamma)}{\sum_\ell \exp(\alpha_\ell - \gamma)}$$

- Now, consider we have N training vectors F_k, each with an associated class label c_k.

Suppose there are N_j training examples of class j and N examples total. Then

$$b_j = \frac{N_j}{N} \longrightarrow \underbrace{b_j = \frac{N_j + \beta}{N + k\beta}}_{\text{regularization}}$$
, where β is some constant and k is the num of classes

Suppose that class j has N_{ij} examples for which the i^{th} feature is 1. Then

$$a_{ij} = \frac{N_{ij}}{N_j} \longrightarrow a_{ij} = \frac{N_{ij} + \alpha}{\underbrace{N_j + 2\alpha}_{\text{Regularization}}}, \text{ for some small value } \alpha$$

- E.g. Suppose we observe N examples of class 0 and M examples of class 1, what is the probability of observing class 0?

Soln:

$$\prod_i P(C_i = j) = \left(\prod_{i: C_i = 0} P(C_i = 0) \right) \left(\prod_{i: C_i = 1} P(C_i = 1) \right)$$

$$= b_0^N \cdot b_1^M$$

$$= b_0^N (1 - b_0)^M$$

$$L(b_0) = N \ln(b_0) + M \ln(1 - b_0)$$

$$\frac{\partial L}{\partial b_0} = \frac{N}{b_0} - \frac{M}{1 - b_0} = 0$$

$$0 = N(1 - b_0) - M b_0$$

$$= N - N b_0 - M b_0$$

$$-N = -N b_0 - M b_0$$

$$N = N b_0 + M b_0$$

$$b_0^* = \frac{N}{N + M}$$

Pros and Cons:

1. Pros
   - Works fast due to the conditional independence assumptions.
   - Works well with high-dimensional data

2. Cons
   - The assumptions may not be easy to satisfy.