NBC 과제

아래와 같은 구조를 가진 자료를 가지고 나이브 베이지안 알고리즘으로 아내가 일하고 있는 가를 분류하기 위해서 Binary classifier을 만든다고 하자.

자료 설명

1. Wife's now working? (binary) 0=Yes, 1=No
2. Wife's age (numerical)
3. Wife's education (categorical) 1=low, 2, 3, 4=high
4. Husband's education (categorical) 1=low, 2, 3, 4=high
5. Number of children ever born (numerical)
6. Husband's occupation (categorical) 1, 2, 3, 4
7. Standard-of-living index (categorical) 1=low, 2, 3, 4=high

---

(자료의 형태는 그대로 보전된다고 가정하고, 연속형 변수는 정규 분포를 따른다고 하자 그리고 범주형 변수의 level은 보전된다.)
그런데 우리는 시점(t_k)마다 데이터 n_k개만 얻을 수 있다. 즉, 우리는 일단 그 자료들만을 가지고 분류기를 만든다. 다음 시점(t_{k+1})에서 우리는 데이터 n_{k+1}개를 새로 얻을 수 있는데, 이 때는 이미 저장 공간 문제로 전 시점에 얻은 데이터 n_k개는 가지고 있지 않는다. 다시 말해 누적된 데이터를 전부 이용해서 분류기를 만드는 방법은 사용할 수 없다. 이를 보완하기 위해 아래와 같은 방법을 생각했다.

0. 초기 시점에서의 데이터로 분류기의 모수들을 구한다. (초기화)
1. 새로운 데이터가 $t(\geq 2)$ 시점에서 얻어진 다면 각 모수들을 업데이트 한다. (update function)
2. 각 시점에서 모수가 업데이트된 분류기는 현재의 모수를 이용해서 새로운 데이터에 대해서 predict 해낼 수 있다. (predict function)

work1.txt, work2.txt, work3.txt, test.txt를 이용해서 위와 같은 형태 나이브 베이지안 분류기를 구현하세요. work1,work2,work3은 각각의 시점 자료입니다. 모수 추정의 경우 연속형에서의 평균, 범주형에서 각 범주의 확률과 prior의 확률입니다. update 함수에 각각의 자료를 따로 넣어도 문제없이 모수가 추정되어야 합니다.