**Mixture Density Networks**

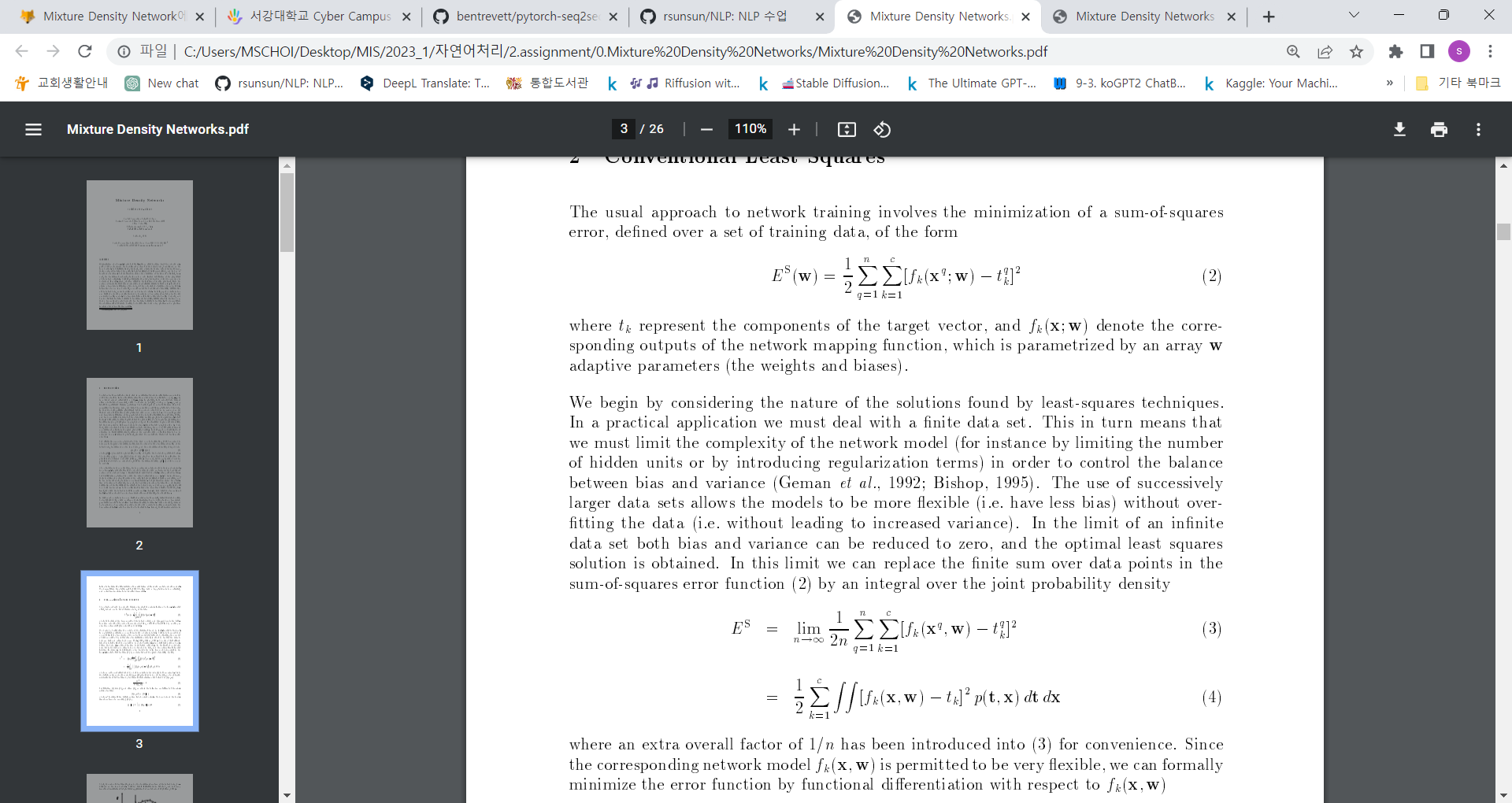
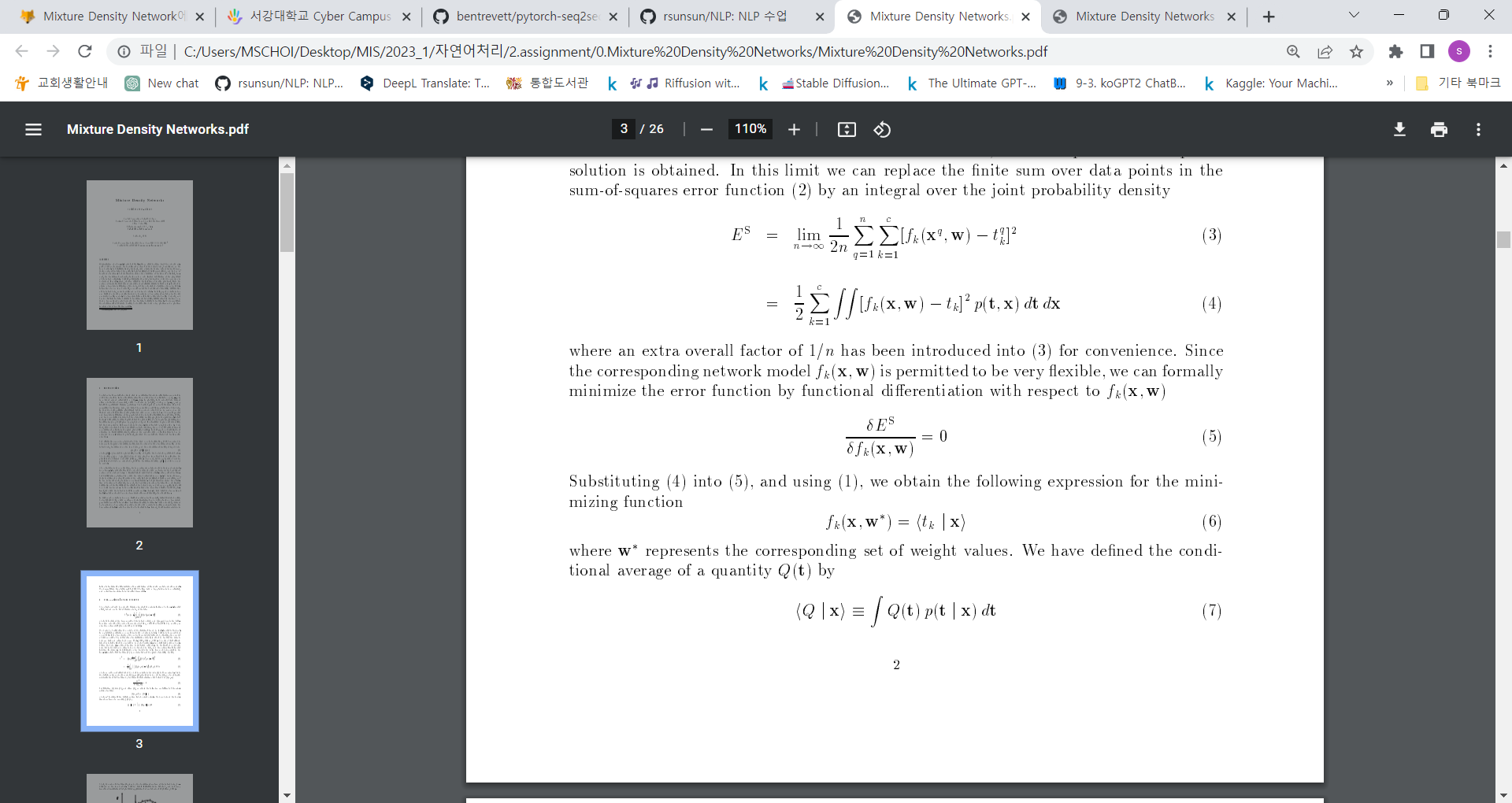
320220065 최미선

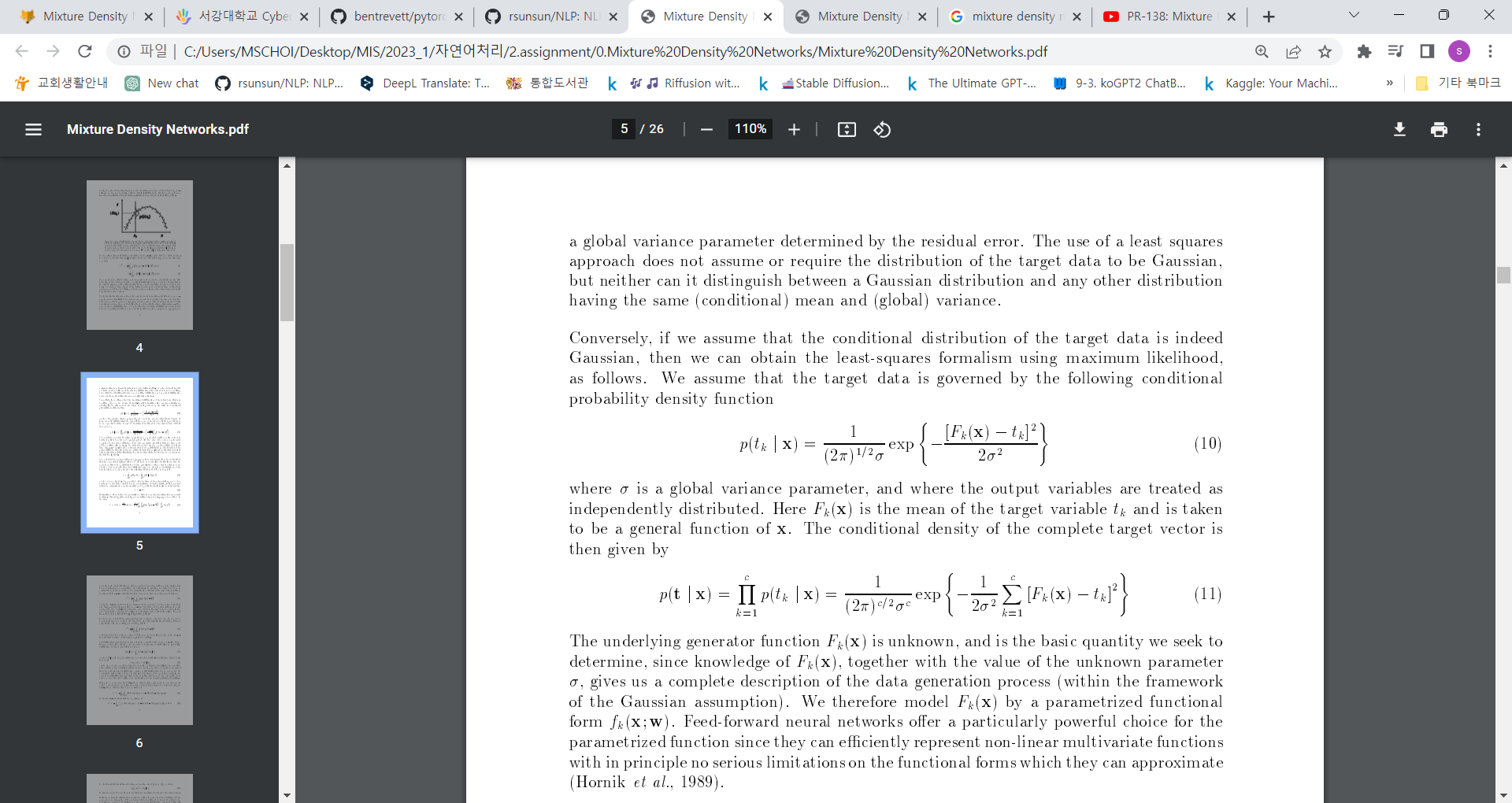
서론

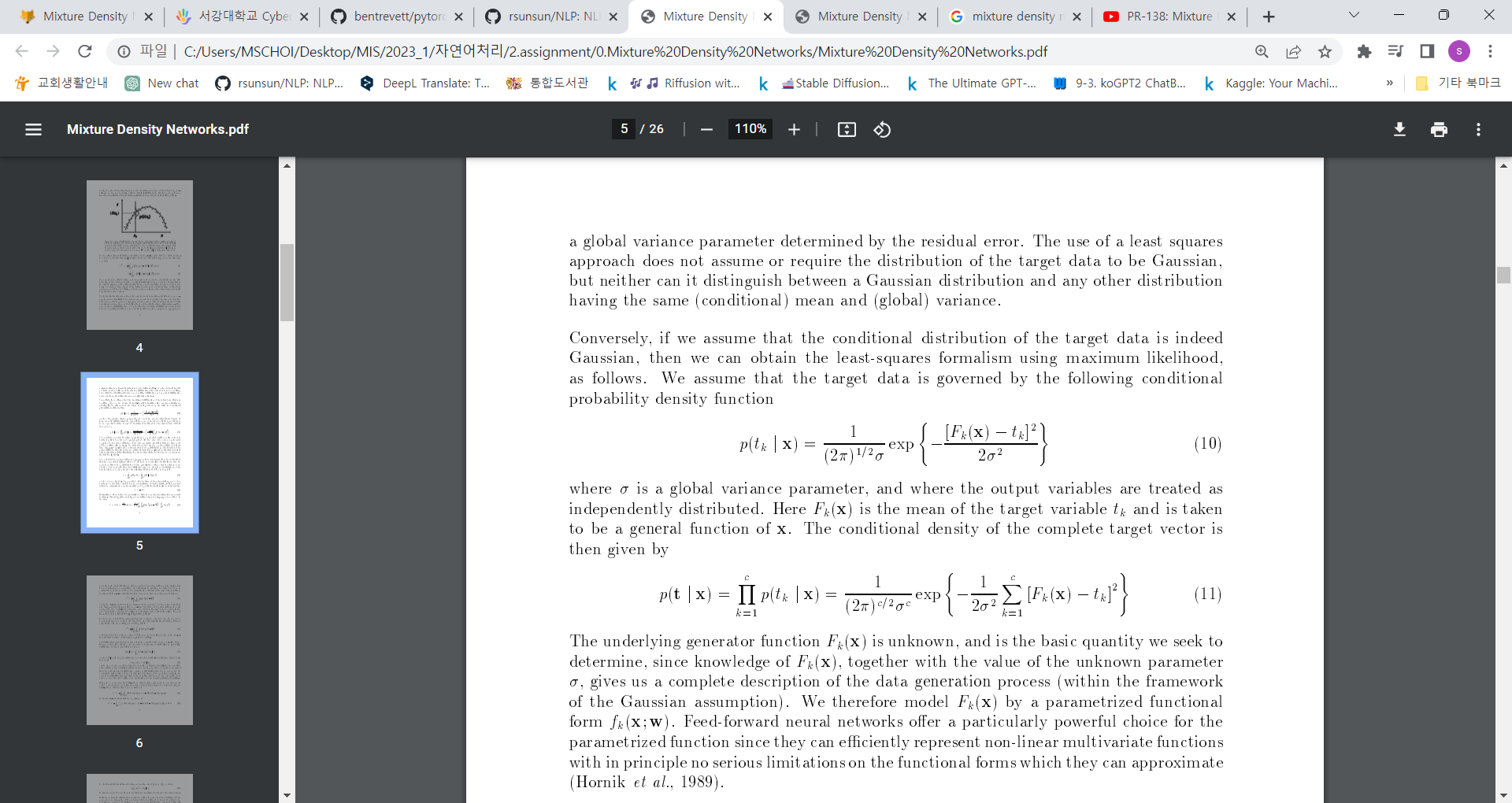
신경망 모델의 네트워크 훈련의 핵심목표는 훈련을 통해 신규 input인 x에 대해 output vector t를 최적으로 예측하는 것입니다. 예측을 하는 모델은 input x와 target t의 결합 확률 밀도 p(x, t)표현될 수 있으며 확률 밀도 함수는 p(x, t) = p(t | x)p(x) 로 표현할 수 있습니다. 이는 input x에 대한 조건부 확률밀도 p(t | x)와 input x의 무조건적인 확률밀도 p(x)의 곱과 같습니다.

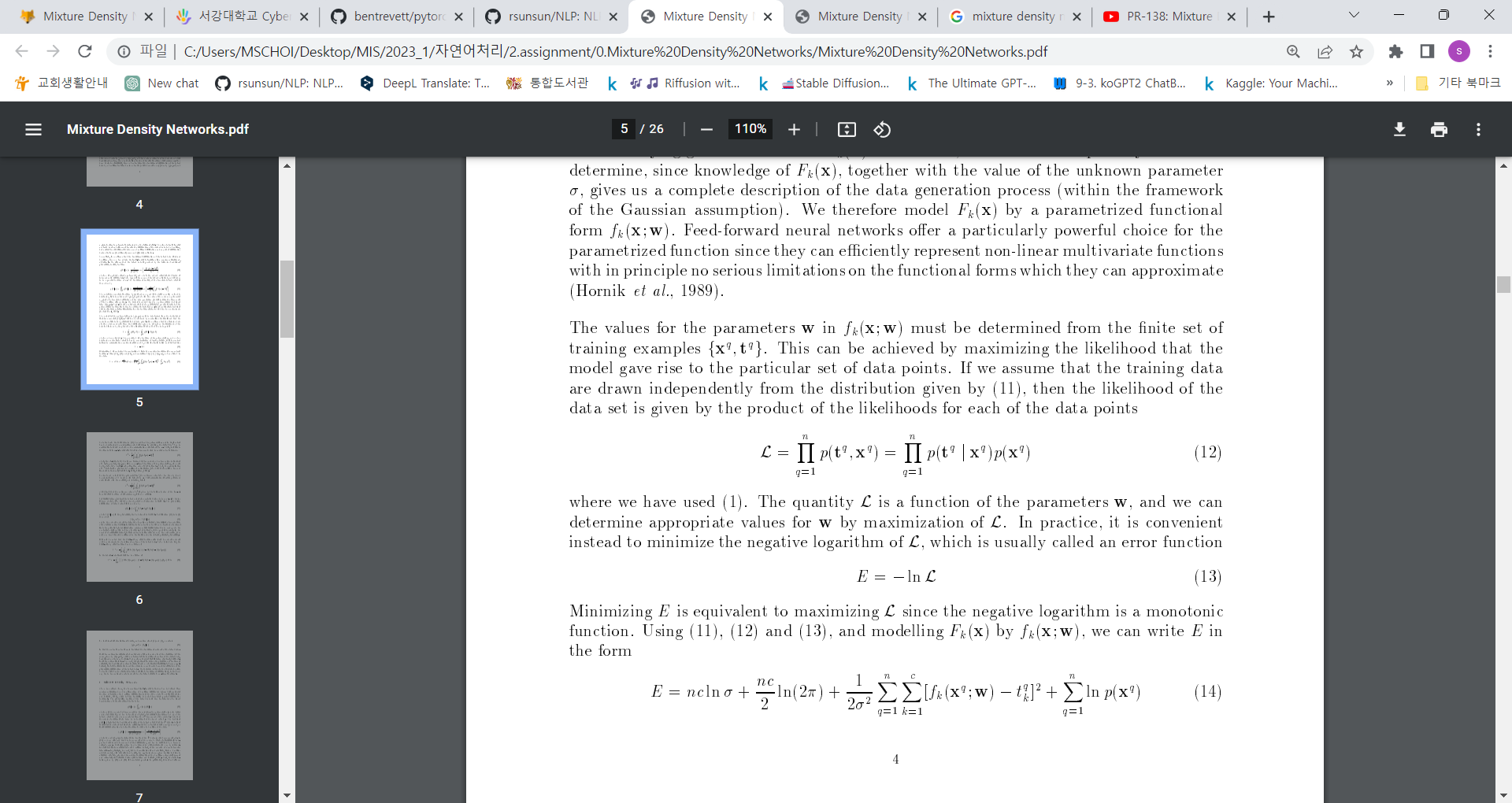
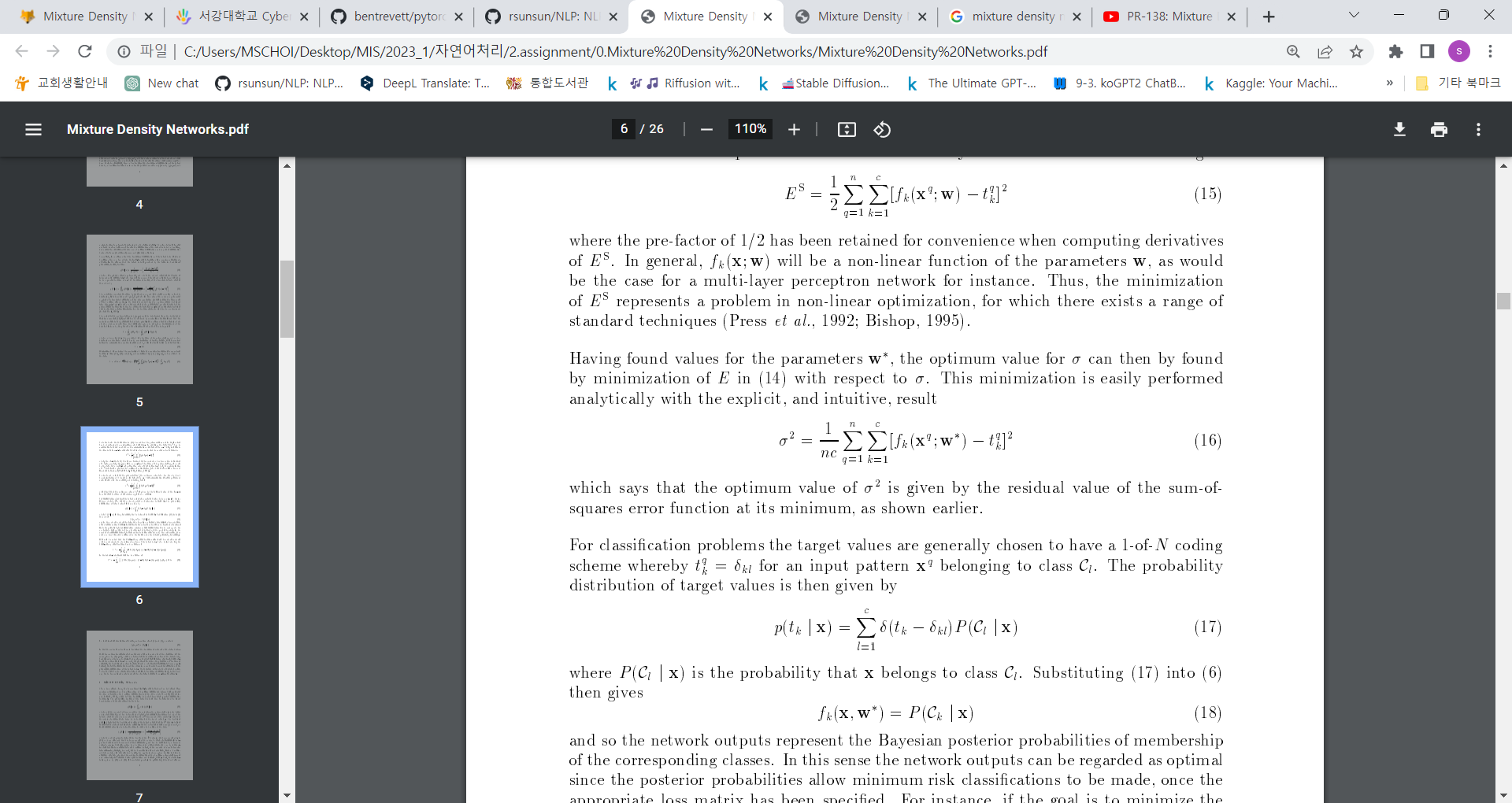
일반적으로 신경망에서 분류 문제 시 최소자승법이나 entropy error 함수를 최소화하여 나온 output 값은 input vector에 조건부로 대응되는 target 데이터의 조건부 평균값을 근사하여 최적의 확률로 나옵니다. 그러나 연속 변수를 예측하는 문제에서는 조건부 평균값은 target 속성을 제한적으로만 설명합니다. 그래서 새로운 input vector에 해당하는 output을 예측하기 위해 다시 한 번 input vector에 조건부로 대응되는 target 데이터의 조건부 확률 분포를 모델링해야 합니다. 즉 일반적인 신경망과 혼합 밀도 모델을 결합하여 새로운 Mixture Density Network(MDN)을 제시하였고 이는 일반적인 신경망이 임의의 함수를 나타낼 수 있는 것과 마찬가지로 임의의 조건부 확률 분포를 표현할 수 있도록 만든 모델이며 로봇 역기구학 문제를 사용하여 검증하였습니다.

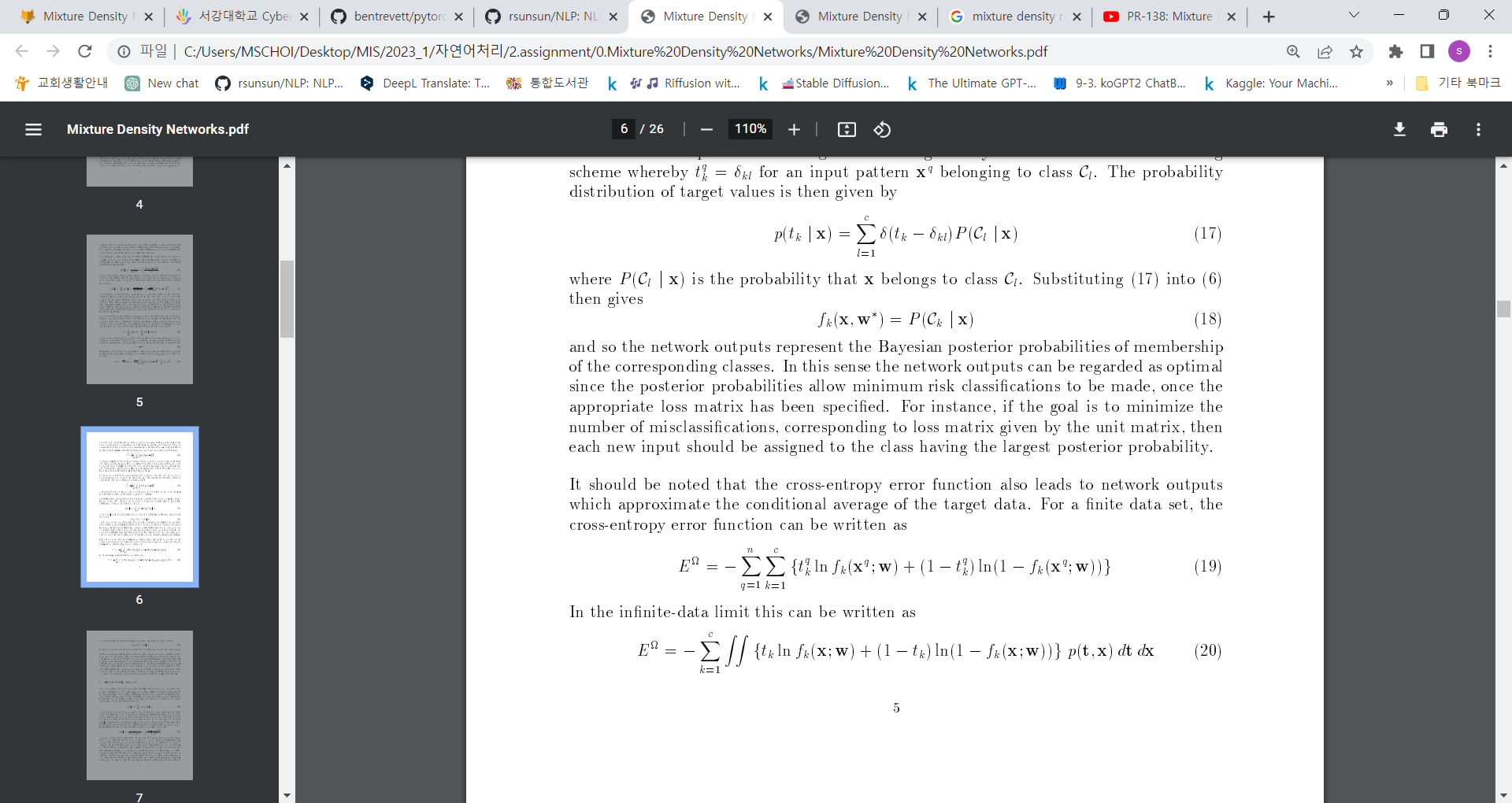
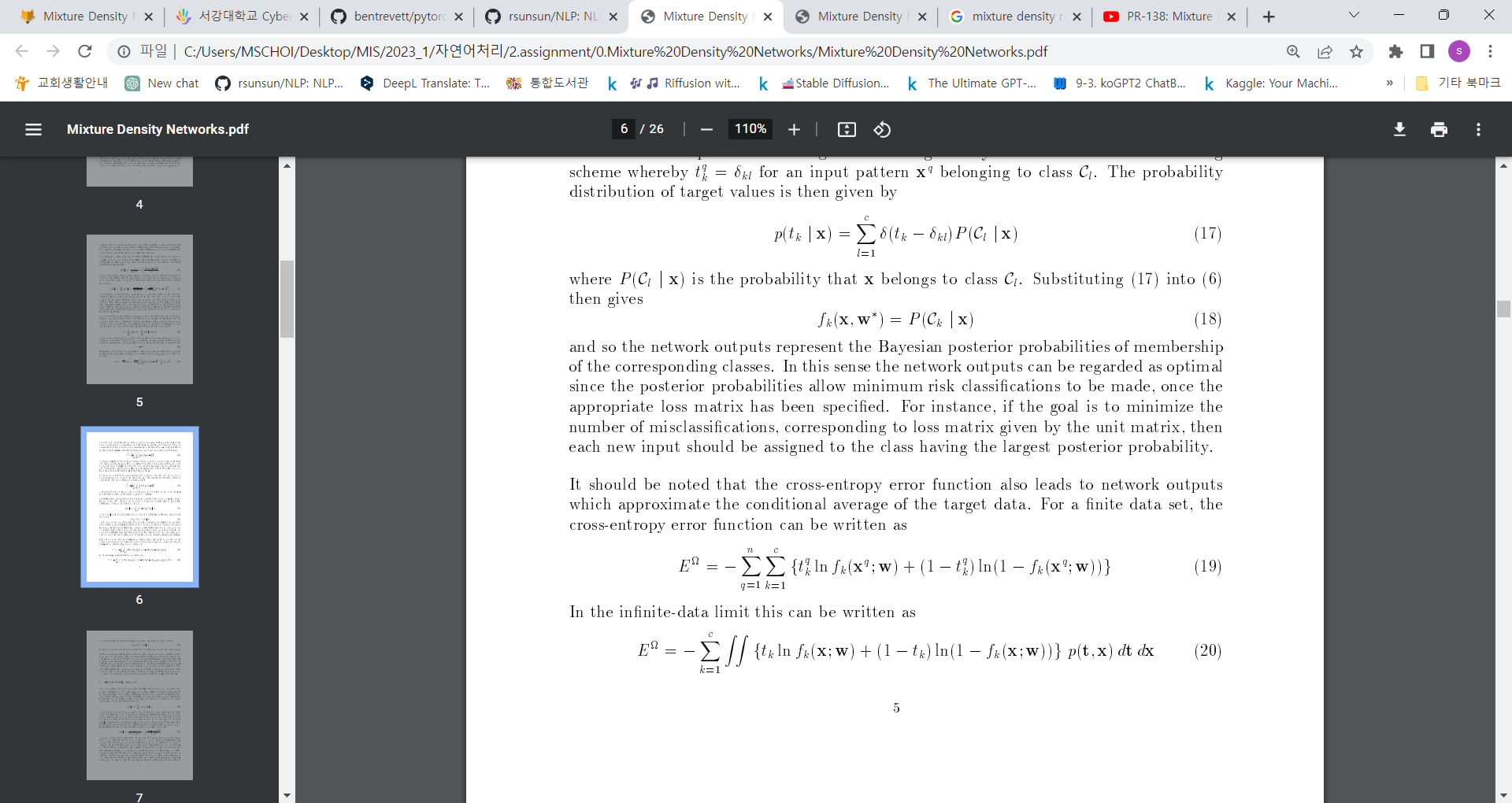
Conventional Least Squares

최소자승법(Conventional Least Squares)은 회귀 분석에서 가장 기본적인 방법 중 하나로, 주어진 입력 변수와 출력 변수 사이의 관계를 모델링하는 함수를 찾는 데 사용됩니다. 이 방법은 주어진 데이터 포인트와 모델 예측 간의 제곱 오차의 합을 최소화하는 함수를 찾는 것입니다. 왼쪽의 식은 신경망의 학습을 위한 비용함수를 나타내는 것으로, 일반적으로 최소제곱법(Least Squares)에서 사용됩니다. 입력 x와 해당 target vector t 사이의 차이의 제곱합을 최소화하도록 신경망의 가중치와 편향을 조절해가며 학습을 수행합니다. 이때 f(x; w)는 신경망의 출력을 나타내며, w는 가중치와 편향의 조합으로 이루어진 매개변수들의 배열입니다. 즉, 최소제곱법은 입력과 출력간의 차이를 최소화하여 모델을 학습하는 방법입니다. 그리고 데이터 집합이 무한대로 가면 바이어스와 분산을 모두 0으로 줄일 수 있으며, 최적의 least squares 해결책을 얻을 수 있고 이는 위의 식 데이터 포인트에 대한 유한 합 대신 결합 확률 밀도 함수에 대한 적분으로 대체할 수 있습니다.

최소 제곱법을 사용하여 일반적인 신경망을 학습시키면 입력 벡터 x에 대한 조건부 평균인 f(x; w)와 이 조건부 평균 주변의 데이터의 평균 분산, 즉 잔차 값으로 표현되는 대상 데이터의 두 가지 통계량을 근사합니다. 이 두 가지 통계량을 알면 조건부 평균과 전체 분산 매개 변수가 결정된 가우시안 함수로 대상 데이터의 조건부 분포를 표현할 수 있습니다. 조건부 밀도 함수는 가우시안 분포를 따르며, 평균값을 F(x)로 가지고 x에 대한 일반적인 함수로 가정합니다.

전체 출력 벡터의 조건부 밀도 함수는 개별 조건부 밀도 함수의 곱으로 표현되며, 출력 변수가 독립적으로 분포된다는 가정 하에 전체 출력 벡터에 대한 조건부 밀도 함수를 구할 수 있습니다. F(x)은 알 수 없는 함수이며, 이 함수를 추정하는 것이 우리의 목표입니다. 따라서 Fk(x)를 매개변수화된 함수 f(x; w)로 모델링하고, 신경망을 사용하여 이 함수를 근사화하고 이때 피드포워드 신경망은 비선형 다변량 함수를 효율적으로 나타내기 위해 가장 좋습니다.

위의 식을 통해 각 데이터 포인트의 likelihood의 곱으로 데이터 집합의 likelihood를 구해서 함수 L을 얻을 수 있으며, 이는 파라미터 w의 함수입니다. 실제로 L의 음의 로그를 최소화하는 것이 편리하며, 이는 E = -lnL 일반적으로 loss function이라고 불리며 신경망 훈련에서 흔히 사용되는 표준 제곱 오차 함수가 됩니다.

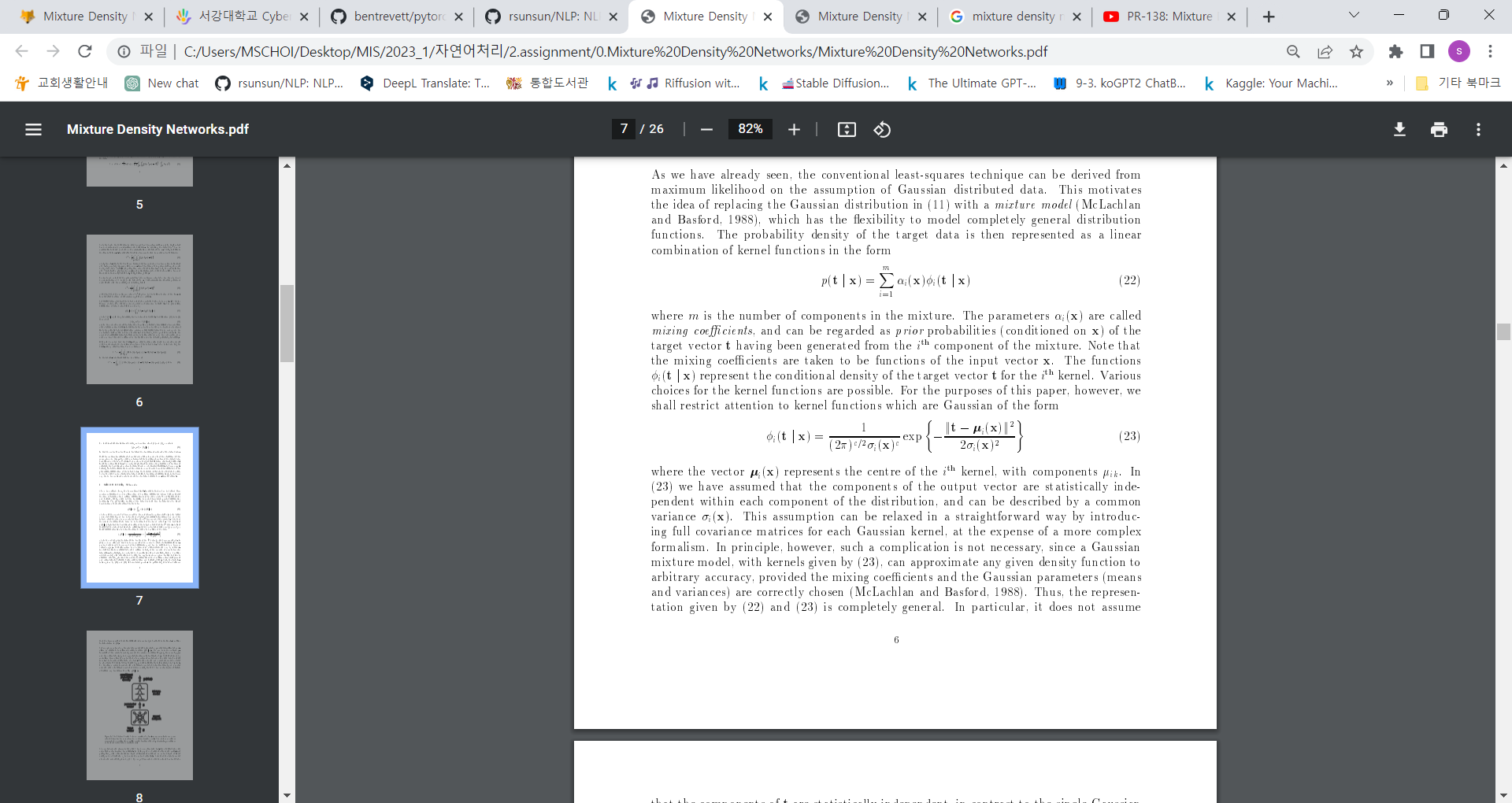
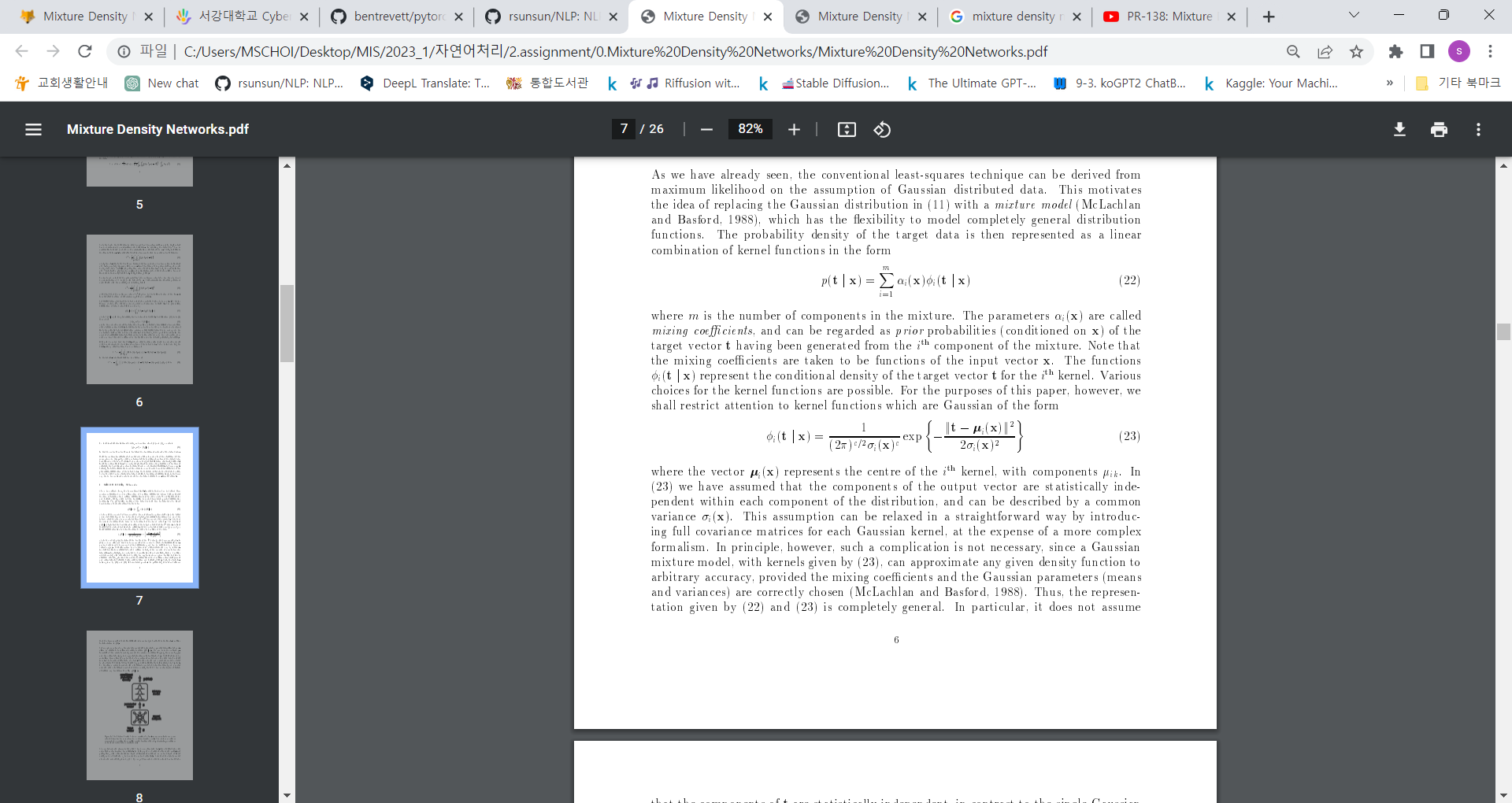
분류 문제의 경우, target 값은 일반적으로 1-of-N 코딩 체계를 가지게 됩니다. target 값의 확률 분포는 옆의 함수 식과 같습니다. 이는 또한 로 표현됩니다.

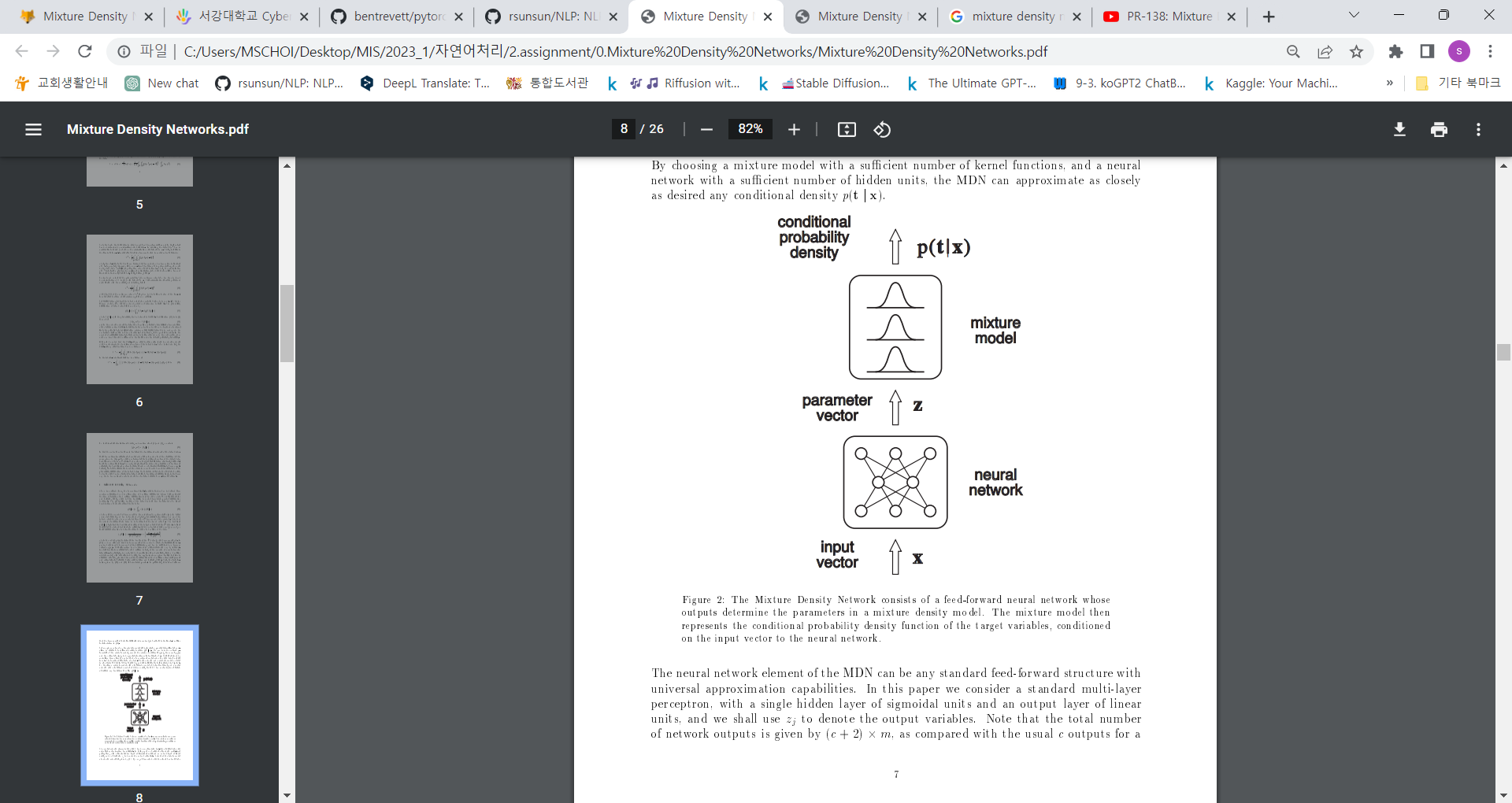
교차 엔트로피 오차 함수도 위와 같이 target 데이터의 조건부 평균을 근사하는 output을 제공합니다.

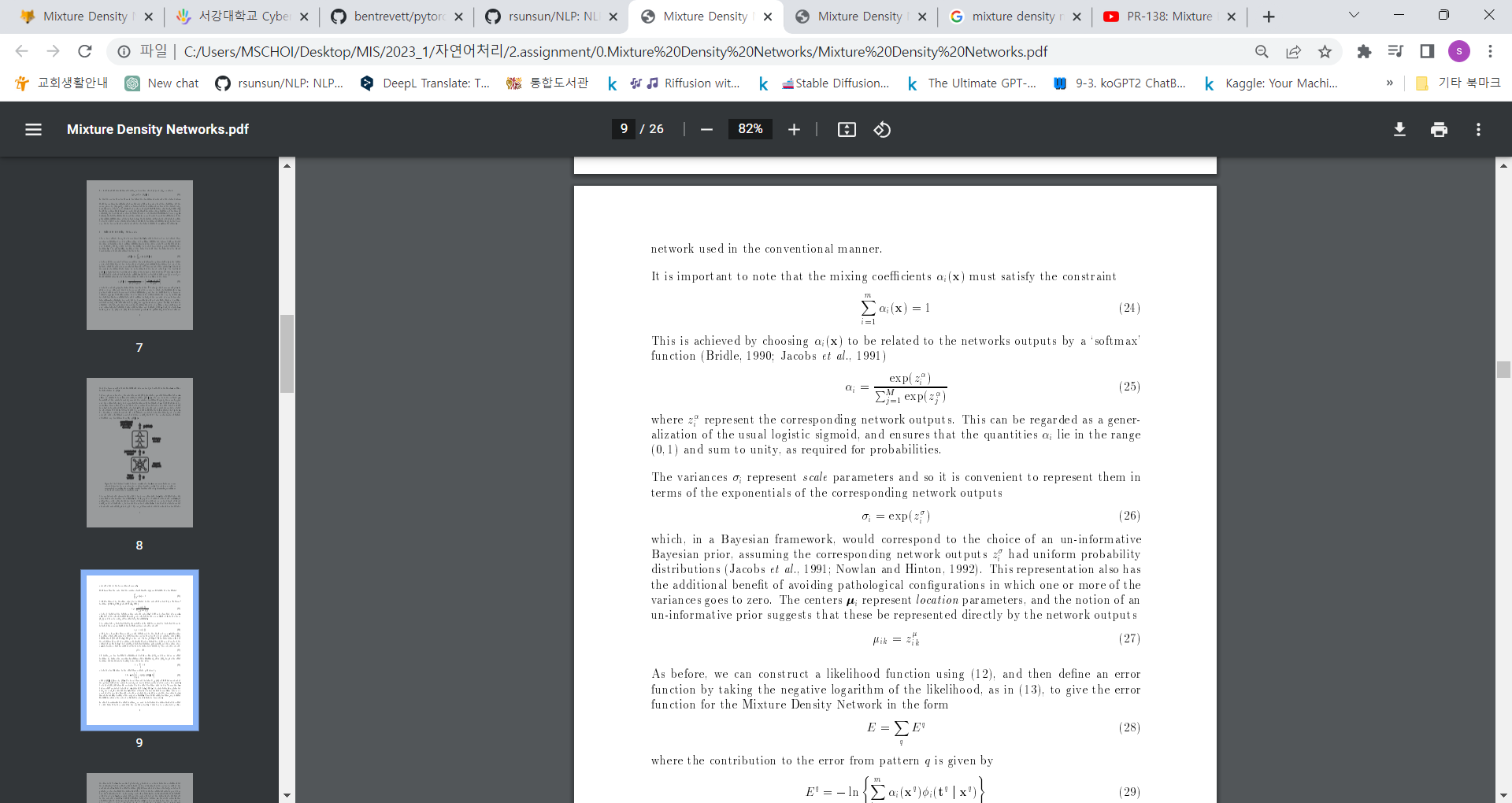
신경망의 대부분의 전형적인 응용에서는 조건부 평균 예측만 사용합니다. 이는 입력 벡터에 대한 조건부 평균을 근사하는 fk(x; w\*)으로 표현되며 이것은 분류 문제에 대한 최적의 해법이었지만 연속 변수의 예측을 다루는 문제에서는 조건부 분포를 나타낼 수 있는 Mixture Density Network이 해법이 될 수 있습니다. 이는 일반적인 신경망이 임의의 비선형 함수를 나타낼 수 있는 것과 같은 방식으로 임의의 조건부 분포를 나타낼 수 있습니다.

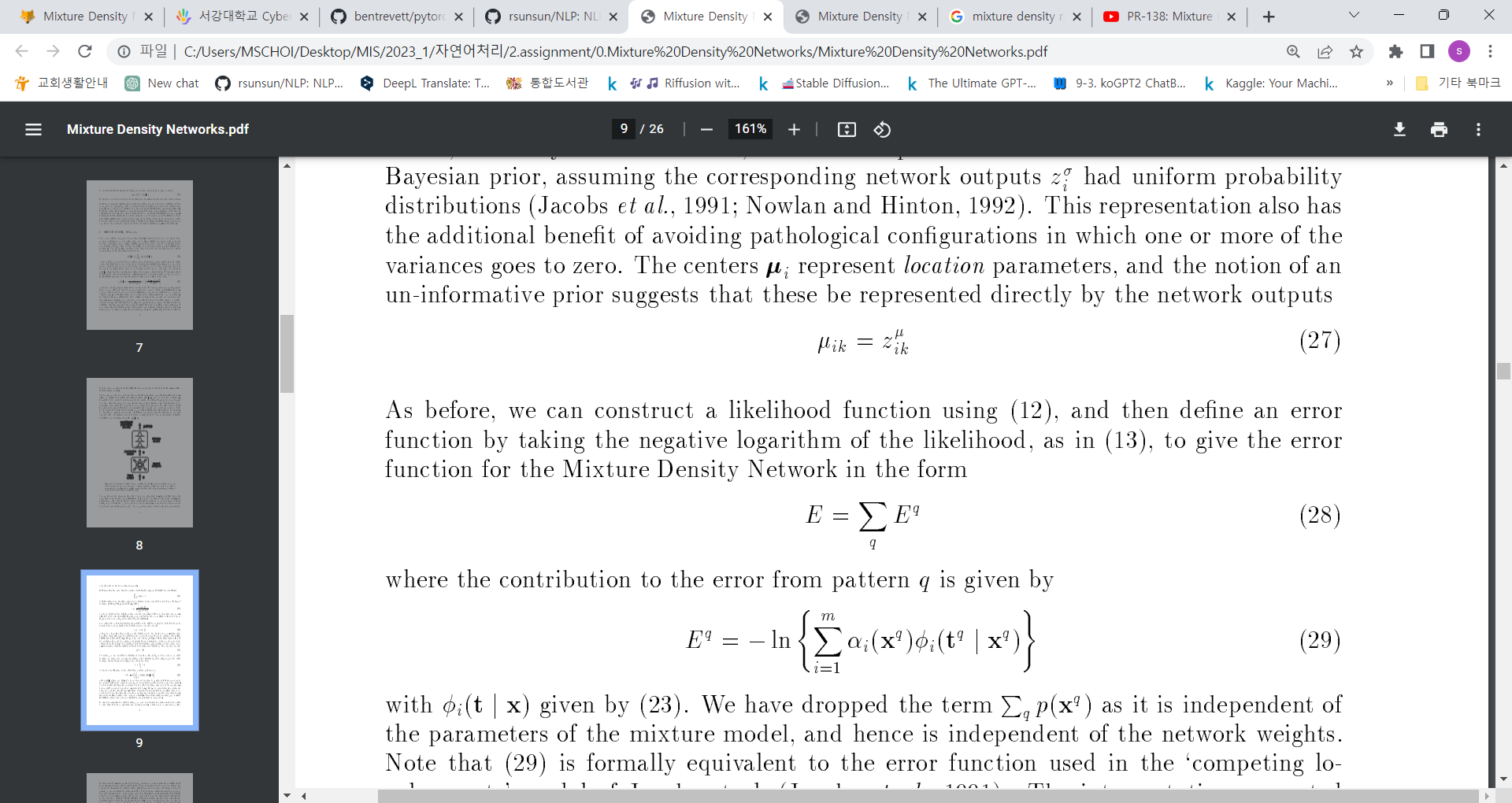
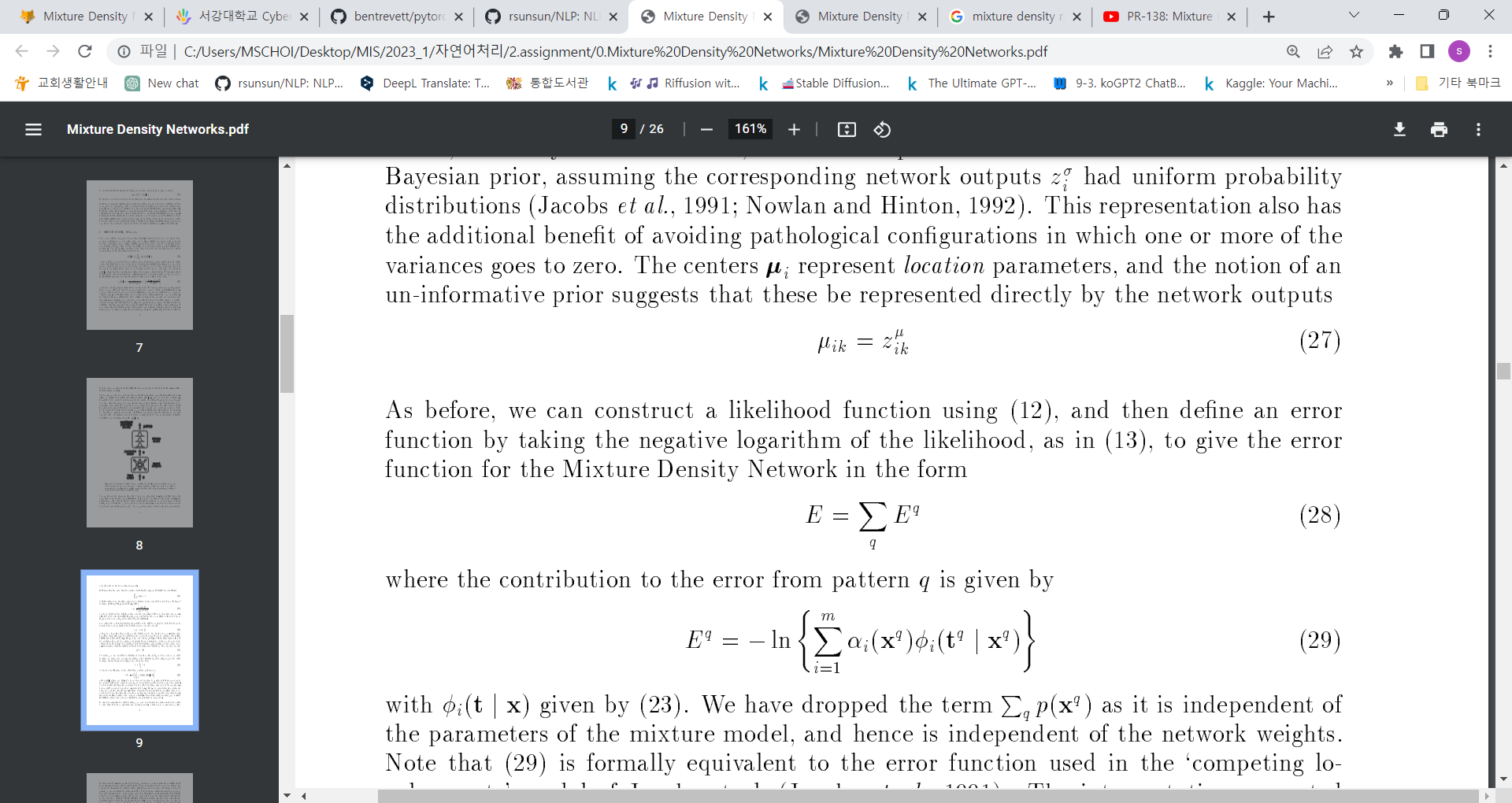
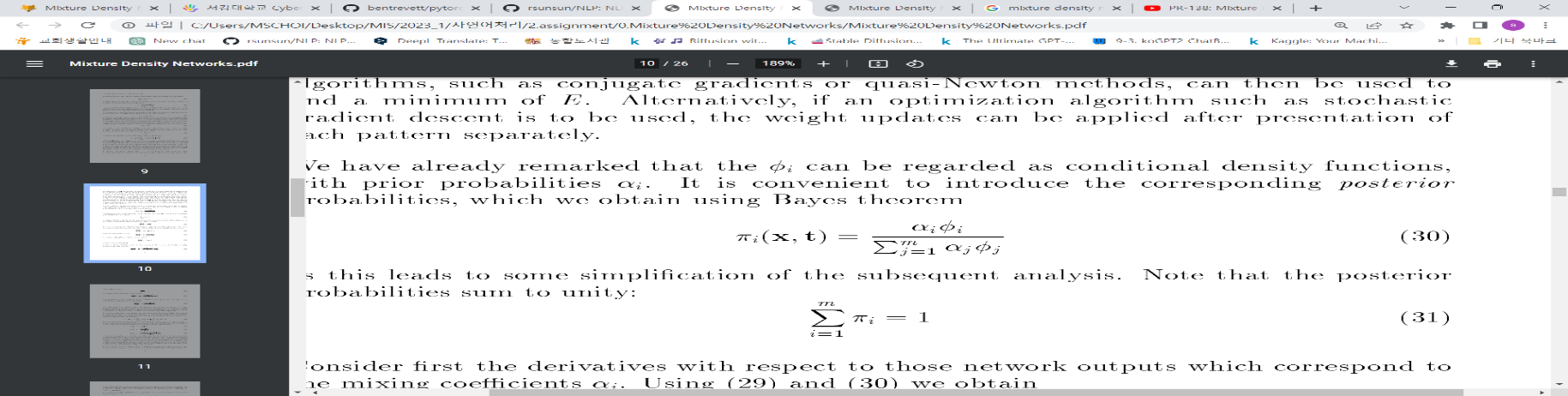
Mixture Density Networks

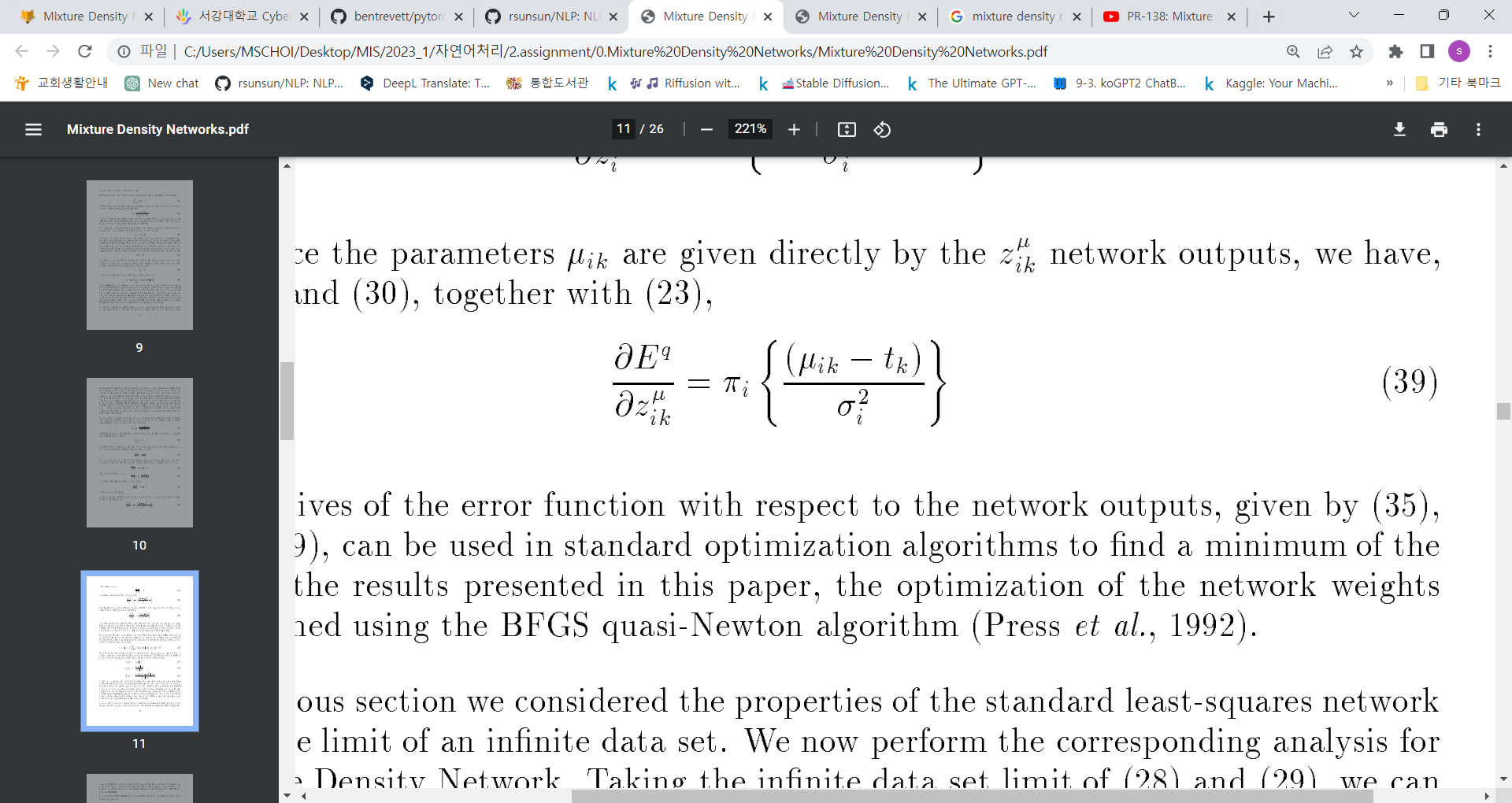
위에서 살펴본대로 최소자승법은 가우시안 분포를 가정한 최대우도법에서 유도될 수 있습니다.이것은 완전히 일반적인 분포 함수를 모델링할 수 있는 유연성을 가진 혼합 모델로 가우시안 분포를 대체하는 아이디어를 제안합니다.

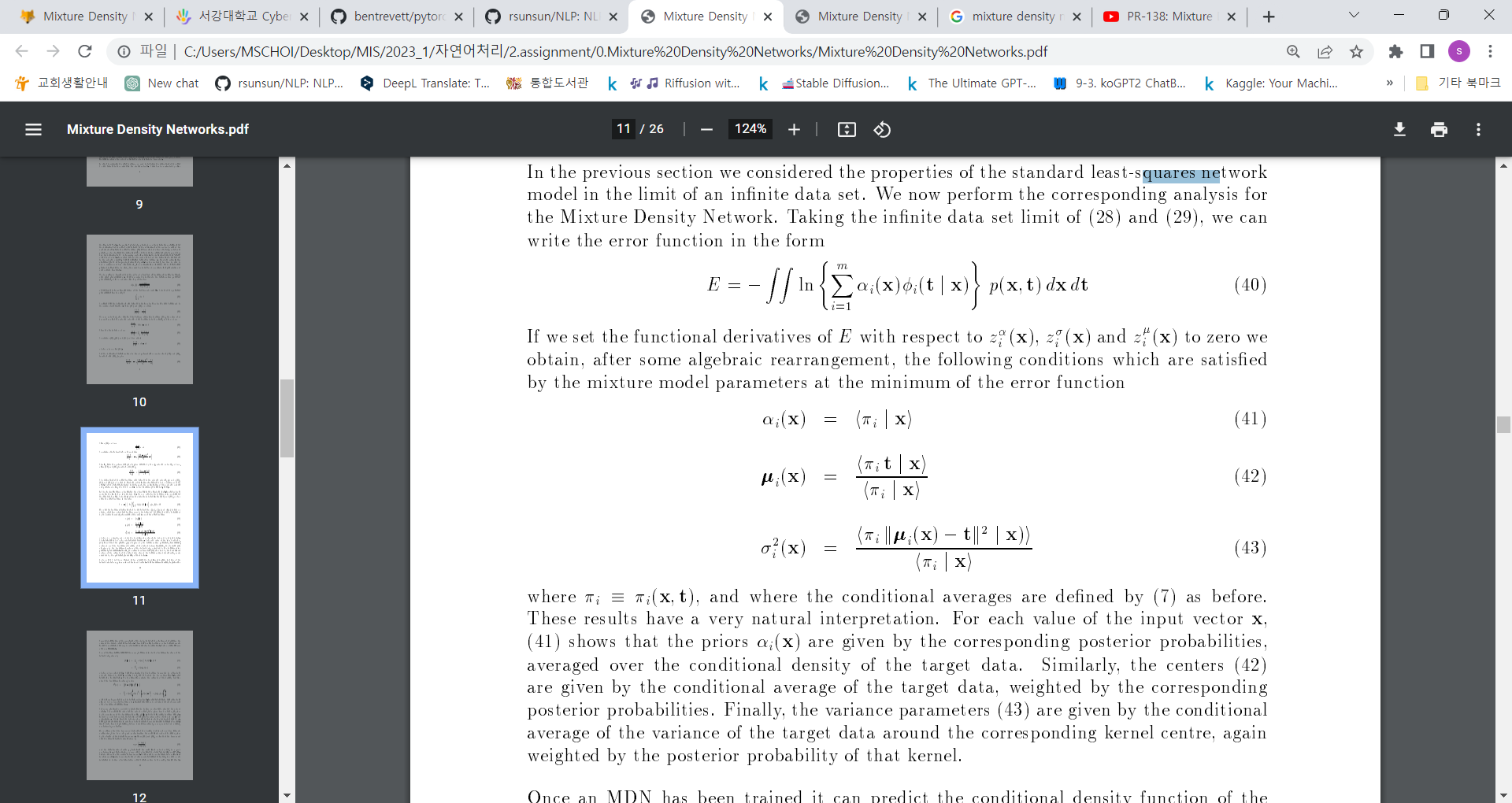
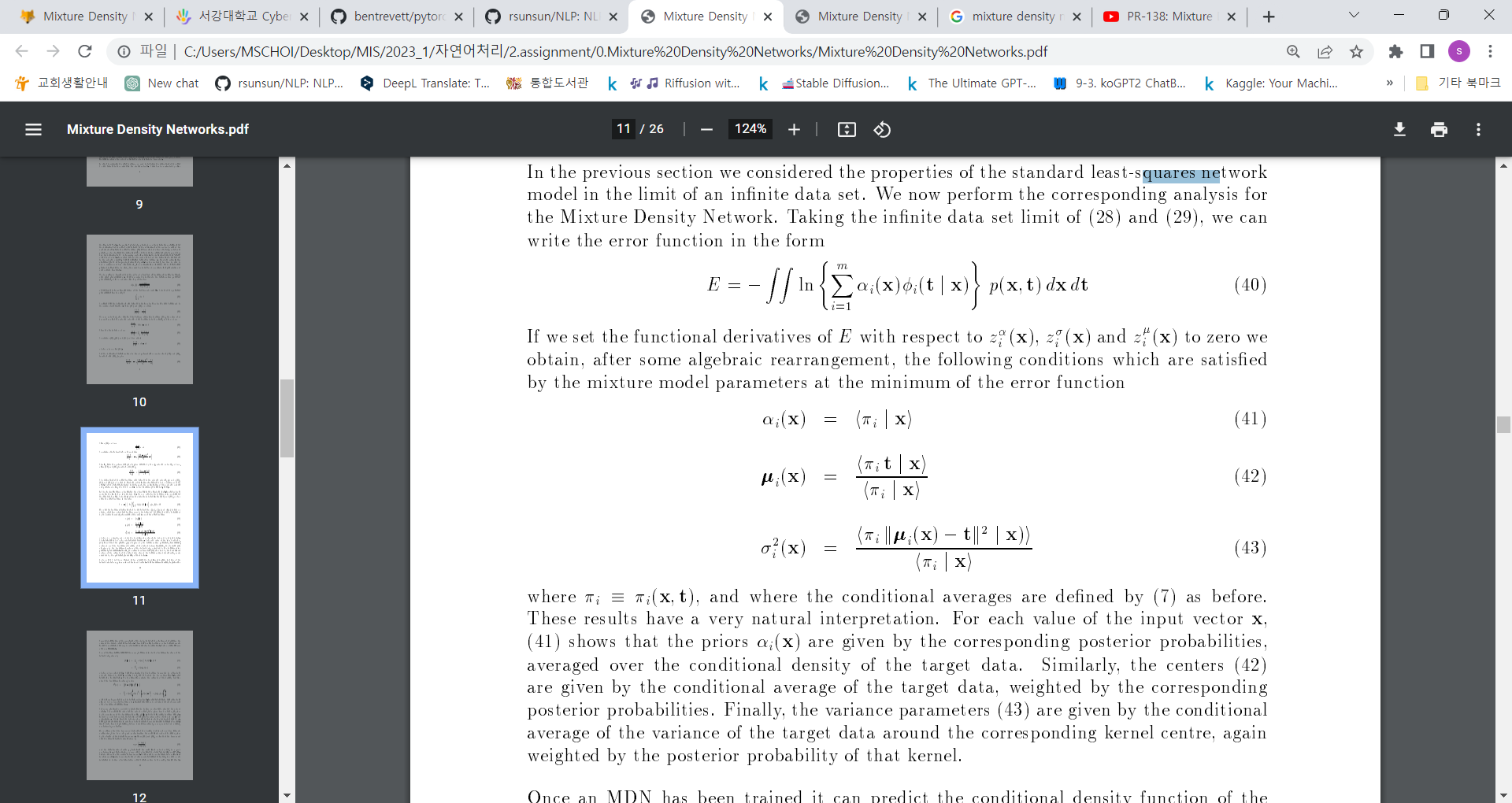
위 식은 혼합 분포(mixture model)를 사용하여 target 데이터의 확률 밀도를 모델링하는 것을 설명하고 있습니다. 이 모델은 여러 개의 커널 함수(µ, kernel function)을 선형 조합으로 결합하여 목표 변수의 조건부 밀도 함수(conditional density function)를 근사하는 방법입니다. 여기서 커널 함수는 가우시안 분포를 따르며, 각 커널 함수마다 중심점과 분산이 다릅니다. 이렇게 여러 개의 가우시안 분포를 결합하면서 목표 데이터의 분포를 모델링하게 됩니다. 가우시안 혼합 모델은 혼합 계수(α*i*(x))와 가우시안 매개 변수 (평균 및 분산)이 올바르게 선택된 경우 임의의 밀도 함수를 임의의 정확도로 근사화 할 수 있어 위 식은 완전히 일반적입니다. 특히 최소자승법에서 다룬 단일 가우시안 표현과 달리 t의 구성 요소가 통계적으로 독립이라는 것을 가정하지 않습니다.

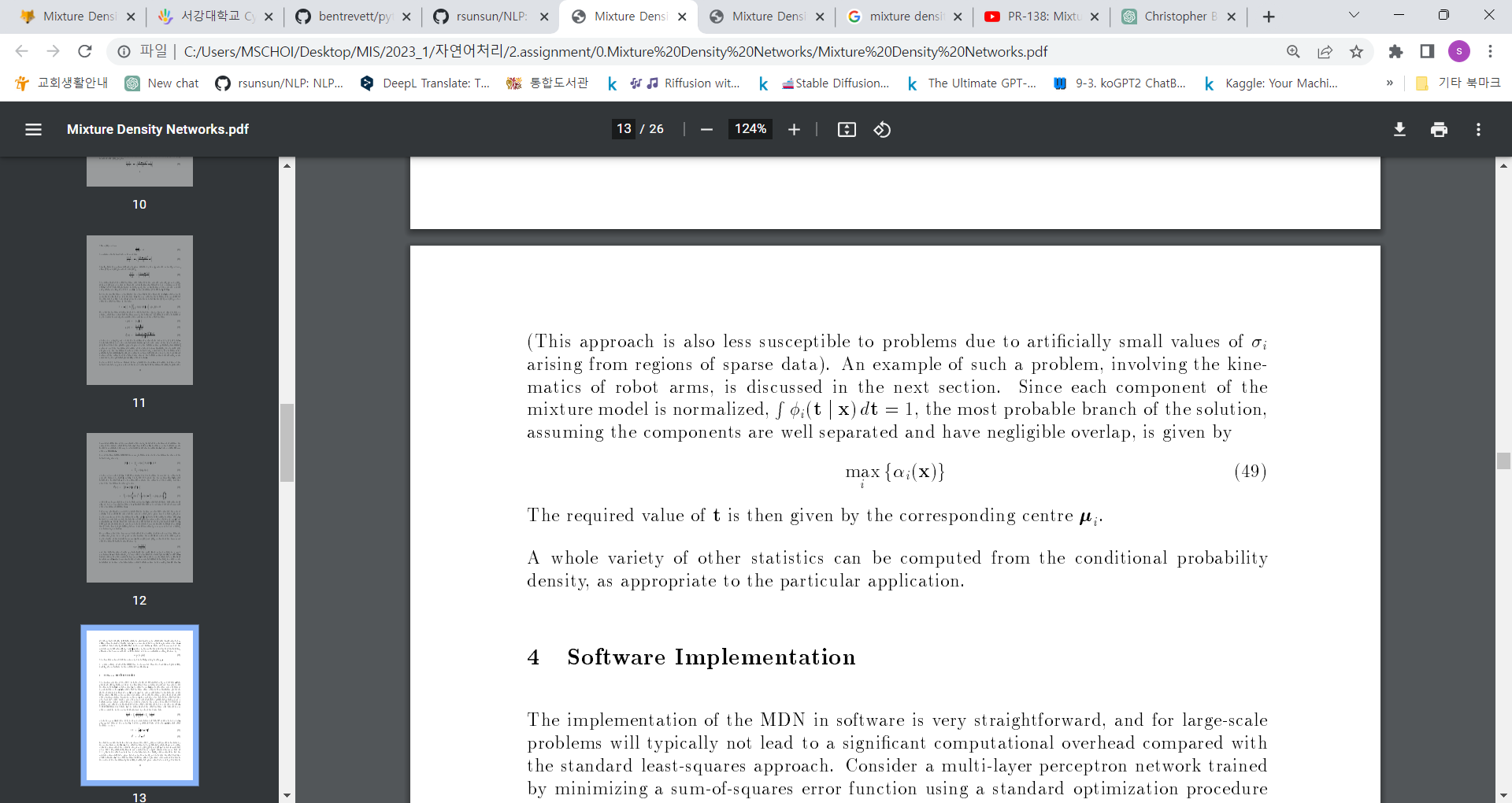
위의 조건부 밀도 함수는 주어진 x값에 대해, 혼합 모형은 임의의 조건부 밀도 함수 p(t|x)를 모델링하는 일반적인 방법을 제공합니다. 혼합 모형의 여러 매개변수, 즉 혼합 계수(α*i*(x)), 평균 µ*i*(x) 및 분산 σ*i*(x)을 일반적인(연속적인) x 함수로 모델링합니다. 이는 x를 입력으로 취하는 전통적인 신경망의 출력을 사용하여 모델링하는 것이며 전방향 신경망과 혼합 모델의 결합 구조를 혼합 밀도 네트워크(MDN)라고 하고 기본 구조는 옆의 그림과 같습니다. 충분한 수의 커널 함수를 가진 혼합 모델과 충분한 수의 은닉 유닛을 가진 신경망을 선택함으로써, MDN은 조건부 밀도 p(t|x)를 원하는 만큼 가깝게 근사화 할 수 있습니다. 혼합 모델은 신경망의 입력 벡터에 조건부로 대상 변수의 조건부 확률 밀도 함수를 나타냅니다.

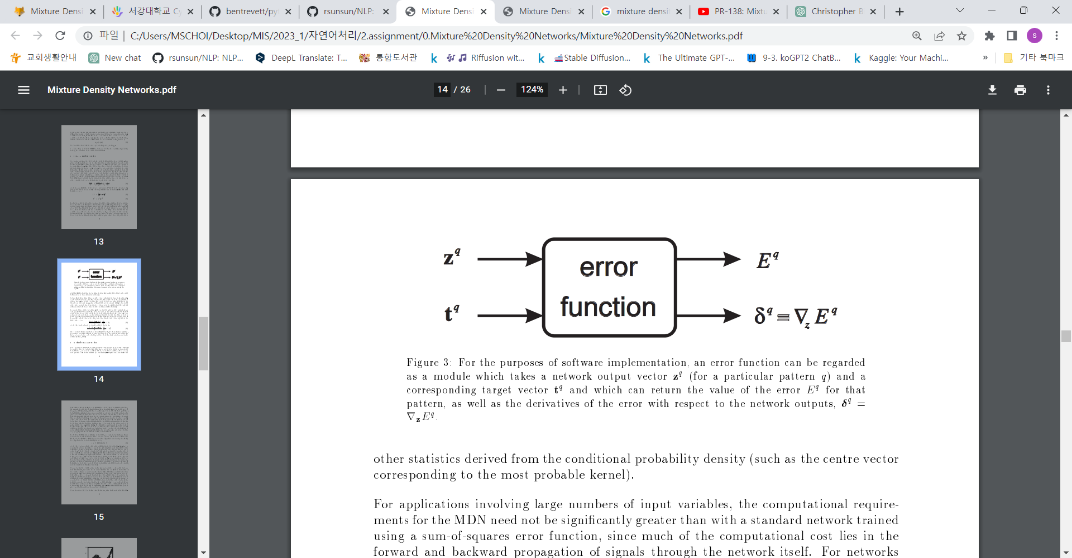
이 논문에서는 시그모이드 유닛의 단일 은닉층과 선형 유닛의 출력층을 가진 표준 다층 퍼셉트론을 고려하며, z*j*로 출력 변수를 나타냅니다. 일반적으로 신경망의 output이 c일 반면 이것은 네트워크 outputs은 (c + 2) \* m이 됩니다. 중요한 점은 혼합 계수는 제약 조건을 충족해야는 것입니다. 이는 softmax함수 통해 가능합니다. 이는 일반적인 로지스틱 시그모이드의 일반화로 간주될 수 있으며, 혼합 계수(α*i*)가 (0, 1) 범위 내에 있고 확률을 위해 필요한 총 합이 1이 되도록 보장합니다.

 Mixture Density Network의 오차 함수는 likelihood함수 및 음의 로그 likelihood를 취하여 정의할 수 있습니다. 에러 함수를 최소화하기 위해서는 신경망의 가중치에 대한 에러 E의 derivative을 계산해야 합니다. 이는 출력 유닛의 활성화에 대한 에러의 derivative를 구하는 적절한 식을 얻는다면 표준적인 '역전파' 절차를 통해계산할 수 있습니다. 우리는 *Фi*를 조건부 확률 밀도 함수로 생각할 수 있습니다. 이 조건부 확률 밀도 함수는 사전 확률 α*i*를 가지고 있습니다. 이에 대응하는 사후 확률을 도입하면 Bayes 정리를 사용하여 합이 1이 되도록 보정된  을 구할 수 있습니다.

혼합 계수 α*i*에 해당하는 네트워크 출력에 대한 오차 함수의 derivative은 옆의 식과 같습니다. 구한 오차 함수의 도함수는 표준 최적화 알고리즘에서 사용됩니다. 이 논문에서 제시된 결과는 BFGS quasi-Newton 알고리즘을 사용하여 네트워크 가중치를 최적화하였습니다.

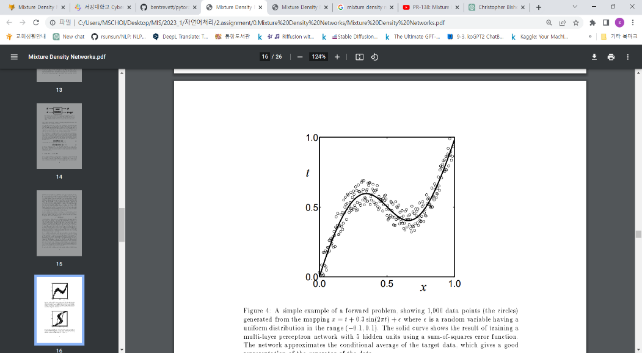
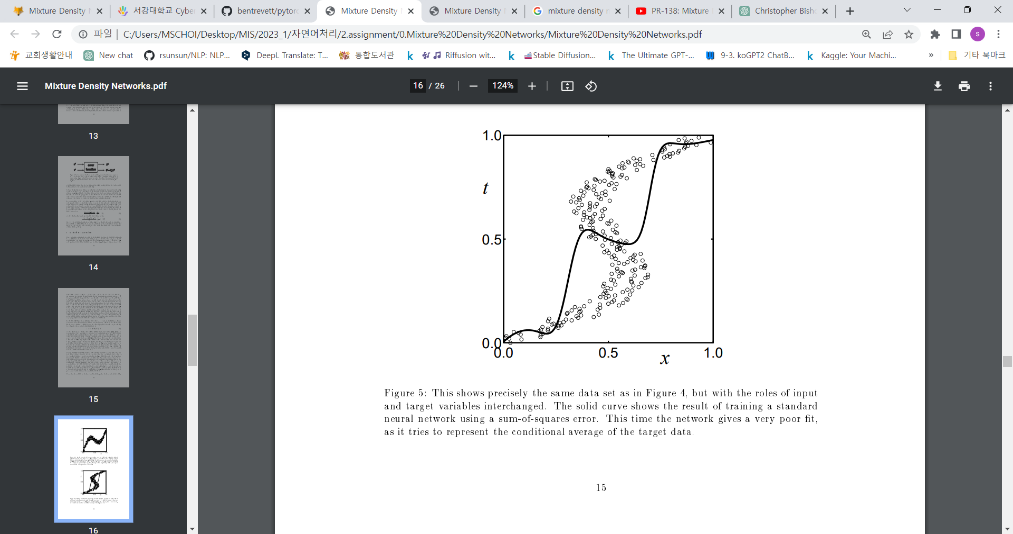
이전까진 무한한 데이터셋에서 표준 최소제곱(standard least-squares) 네트워크 모델의 속성을 고려했듯이 혼합 밀도 네트워크의 무한한 데이터셋 한계를 취하면 옆의 식처럼 오차 함수가 정의됩니다. 그리고 혼합 모델 매개 변수가 충족해야 하는 조건은 옆의 식과 같습니다. 입력 벡터 x의 각 값에 대해 사후 확률 밀도를 통해 가중 평균 된 해당 우선 순위 α*i*(x)가 나타납니다. 마찬가지로 µ*i*(x) 해당 사후 확률로 가중 평균 된 대상 데이터의 조건부 평균으로 주어집니다. 마지막으로 분산 매개 변수 σ*i*(x)는 해당 커널 센터를 중심으로 대상 데이터의 분산의 조건부 평균으로, 다시 해당 k의 사후 확률로 가중치가 부여됩니다. MDN이 학습된 후에는 입력 벡터의 임의의 값에 대해 대상 데이터의 조건부 확률밀도 함수를 예측할 수 있습니다. 이 조건부 확률밀도 함수는 대상 벡터 값을 예측하는 문제에 대한 데이터 생성기의 완전한 설명을 나타냅니다. 이 밀도 함수에서 다른 응용 프로그램에서 유용한 특정 통계량을 계산할 수 있습니다. 본 논문은 조건부 평균에 해당하는 평균값, 조건부 평균값 주변의 밀도 함수의 분산 등을 구하였습니다.

입력 벡터 x가 주어졌을 때, 출력 벡터의 가장 가능성 높은 값을 최대한 찾는 것이 중요합니다. 이 때문에 조건부 밀도함수 p(t|x)의 최대값은 비선형 최적화 문제로 나타납니다. 조건부 밀도 함수의 구성 성분이 서로 겹치지 않을 정도로 충분히 분리되어 있다고 가정하면, 가장 가능성 높은 값 t는 매우 정확하게 가장 높은 구성 성분의 중심으로부터 얻을 수 있습니다. 이것은 한정된 분기가 있는 다중 값 매핑과 관련된 응용 분야에 적합하며, 가장 가능성 높은 분기에 해당하는 대표 벡터를 찾는 데 유용합니다. 혼합 모델의 각 구성 요소는 정규화되므로 ∫*Фi* (t | x) dt = 1이 되며, 구성 요소가 잘 분리되어 있고 겹침이 없다고 가정하면, 해결의 가장 가능성이 높은 가지(branch)는과 같습니다.

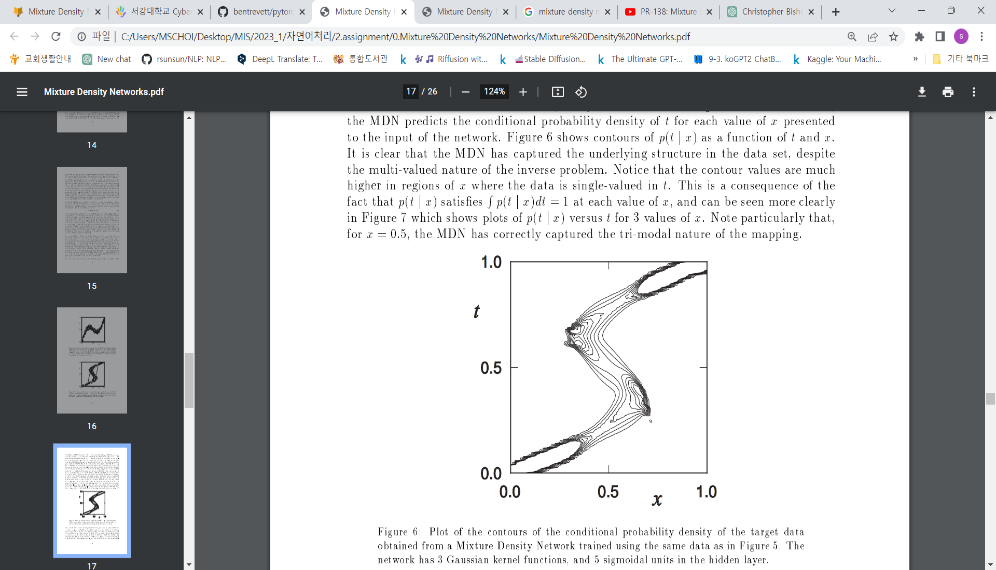
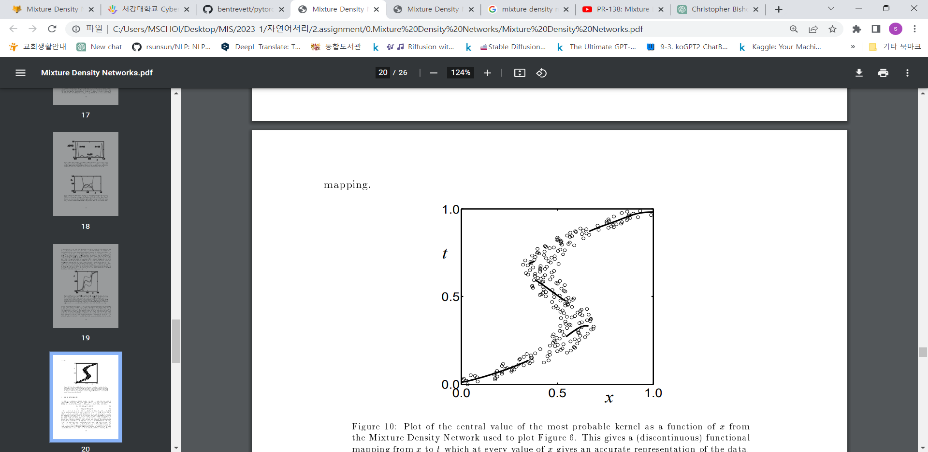
Software Implementation

MDN을 사용하는 경우 입력 변수가 많은 경우에도 일반적인 가중치 학습 방식을 사용하는 네트워크와 비교하여 계산 요구 사항이 크게 증가하지는 않습니다. 이는 네트워크를 통한 신호의 전진 및 역전파에서 대부분의 계산 비용이 발생하기 때문입니다. 따라서 입력 변수가 많은 네트워크의 경우 첫 번째 레이어의 가중치 수가 많아지기 때문에, 이러한 계산 요구 사항은 오차 함수와 그 도함수를 평가하는 비용을 초과할 수 있습니다.

Simple Inverse Problem

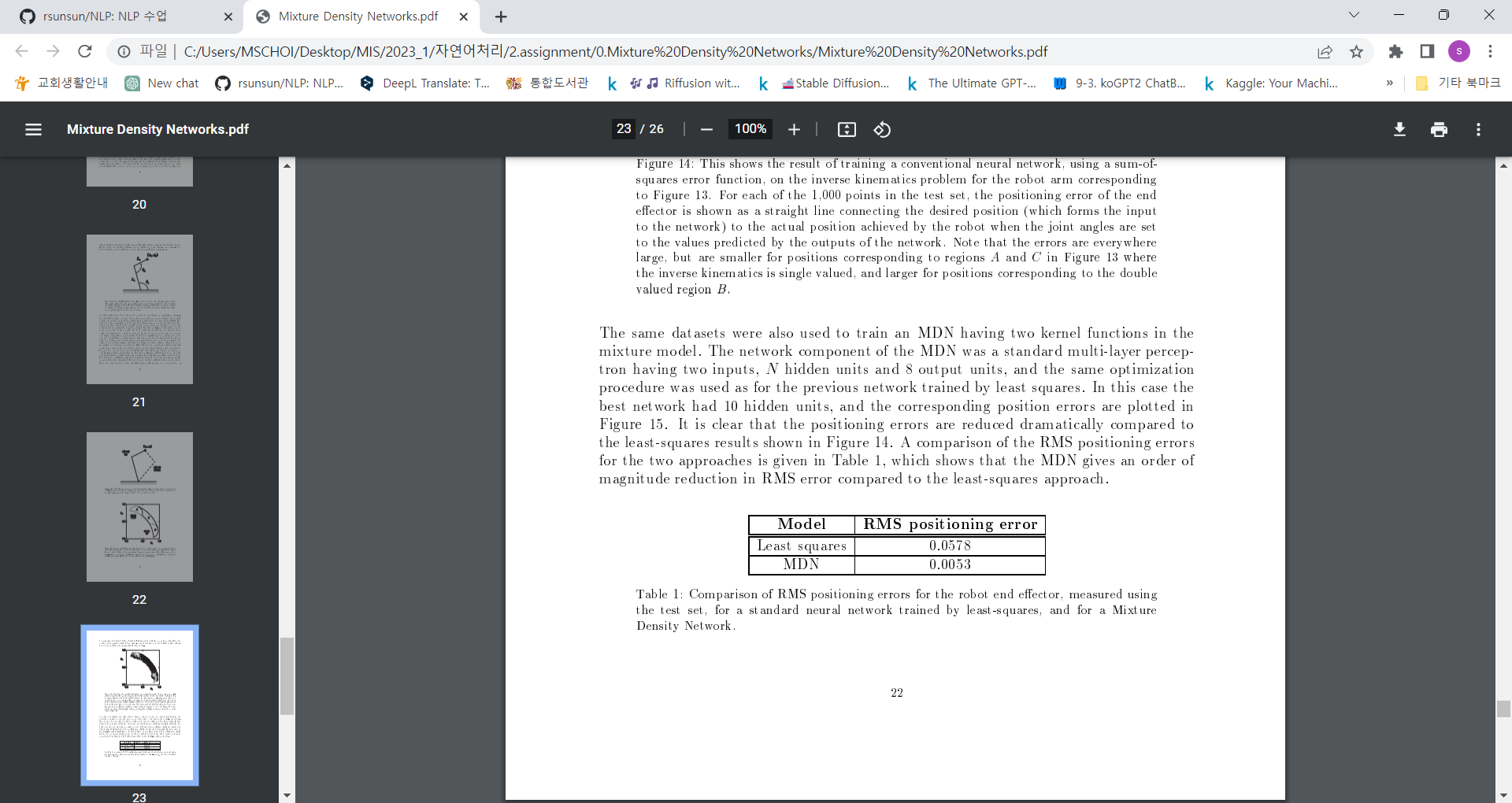
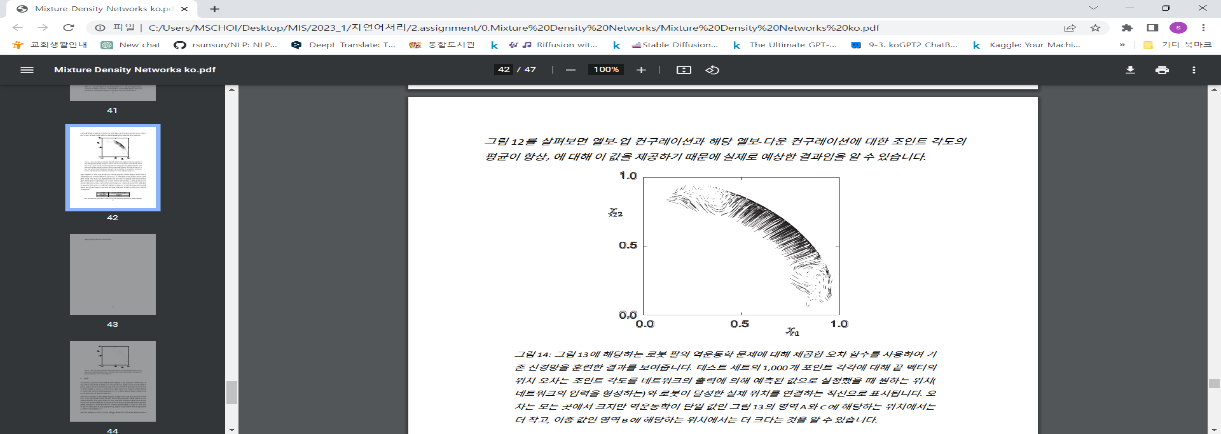
아래 첫번째 그림은 sum-of-squares error function을 사용한 5개의 은닉층(hidden unit)을 가진 multi-layer perceptron 신경망을 학습 시 신경망이 target data의 조건부 평균(conditional average)을 근사하고 있다는 것을 나타내며, 이러한 결과가 데이터 생성 함수(generator of the data)를 잘 표현하고 있다는 것을 보여줍니다.

아래 그림과 같은 데이터셋을 사용하고 입력과 출력 변수의 역할이 바뀐 경우를 보여줍니다. 이번에는 표준 신경망이 제곱 오차를 사용하여 학습되었으며, 신경망은 목표 데이터의 조건부 평균을 나타내려고 하므로 매우 나쁜 적합성을 보입니다. 입력 변수가 출력 변수의 역할을 수행할 때 데이터 생성자를 효과적으로 근사하기 위해 조건부 분포를 모델링하는 MDN을 사용하는 것이 유용할 수 있습니다.

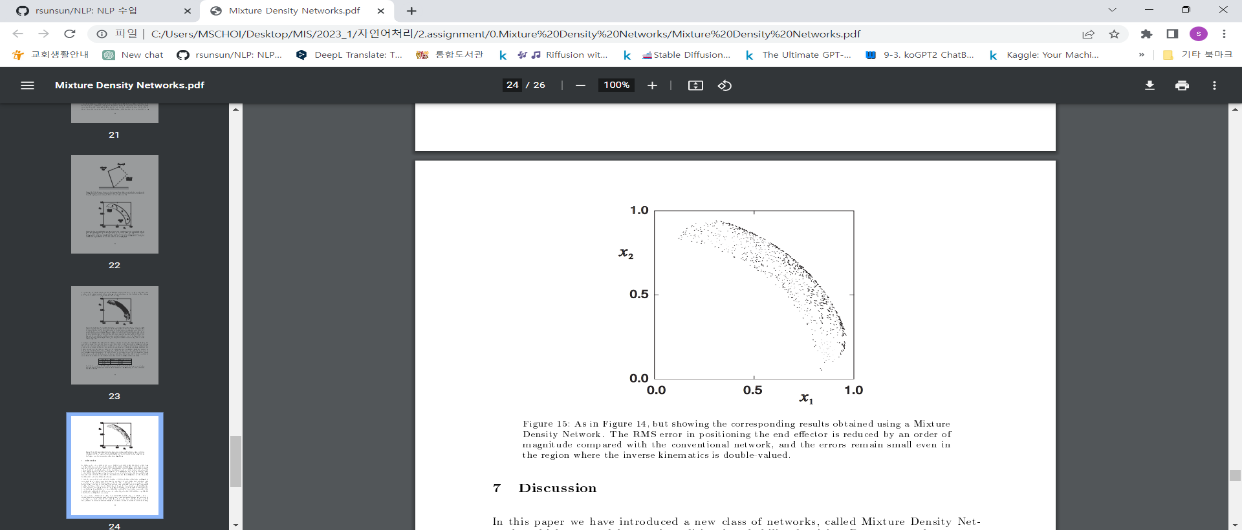
동일한 데이터를 사용하여 훈련된 혼합 밀도 네트워크(Mixture Density Network)에서 얻은 대상 데이터의 조건부 확률 밀도의 등고선(contours)을 보여줍니다. 이 네트워크는 3개의 가우시안 커널 함수와 5개의 시그모이드 유닛을 가진 은닉층(hidden layer)을 가지고 있습니다. 이 그림은 대상 변수(t)의 값에 대한 조건부 밀도 추정치의 등고선을 보여주고 있습니다. 이 그림에서 등고선은 조건부 밀도 함수의 값을 나타냅니다. 더 높은 등고선은 대상 변수(t)의 값에 대한 확률이 높은 지역을 나타내며, 낮은 등고선은 대상 변수(t)의 값에 대한 확률이 낮은 지역을 나타냅니다. 이 등고선 그림을 통해, Mixture Density Network는 입력 변수(x)를 통해 대상 변수(t)를 추정하는 데 유용하다는 것을 알 수 있습니다.

Mixture Density Network의 가장 가능성이 높은 커널의 중심값을 x에 대한 함수로 나타낸 그래프입니다. 이 그래프는 x에서 t로의 불연속 함수 매핑을 제공하며 모든 x 값에서 데이터를 정확하게 나타냅니다. 이 그림은 첫번째 sum-of-squares error function에서 보여진 전통적인 신경망에서 얻은 해당 결과와 비교해 볼 수 있습니다.

Robot Kinematics

혼합 밀도 네트워크의 두 번째 응용 사례로, 그림 11과 같이 간단한 2link 로봇 팔의 운동학을 고려합니다. 테스트 세트와 최소자승으로 훈련된 표준 신경망, 혼합 신경망을 사용하여 측정한 2link 로봇 팔의 운동학의 RMS 위치 오차 비교하였고 MDN은 최소자승 접근법에 비해 최소 제곱 접근 방식에 비해 RMS 오류를 크게 감소시킵니다.

최소자승으로 훈련된 표준 신경망의 오류(RMS)는 영역 A와 C에 해당하는 위치에서는 더 작고 이중 값인 영역 B에 해당하는 위치의 경우 더 큽니다.

MDN의 RMS 오차는 위의 표준 신경망에 비해 크게 감소하며, 역 운동학이 두 배인 영역에서도 오차가 작습니다.