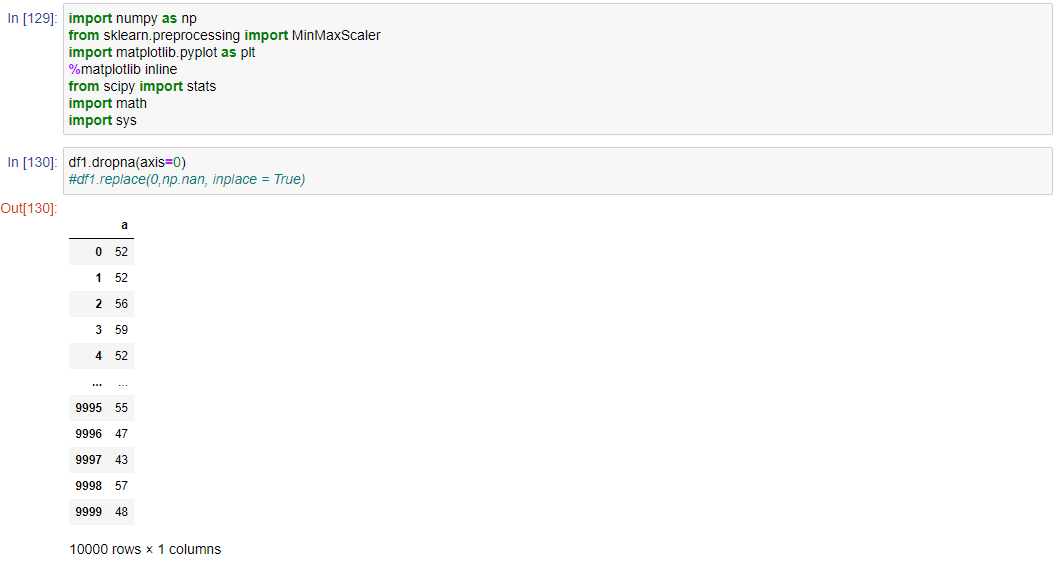
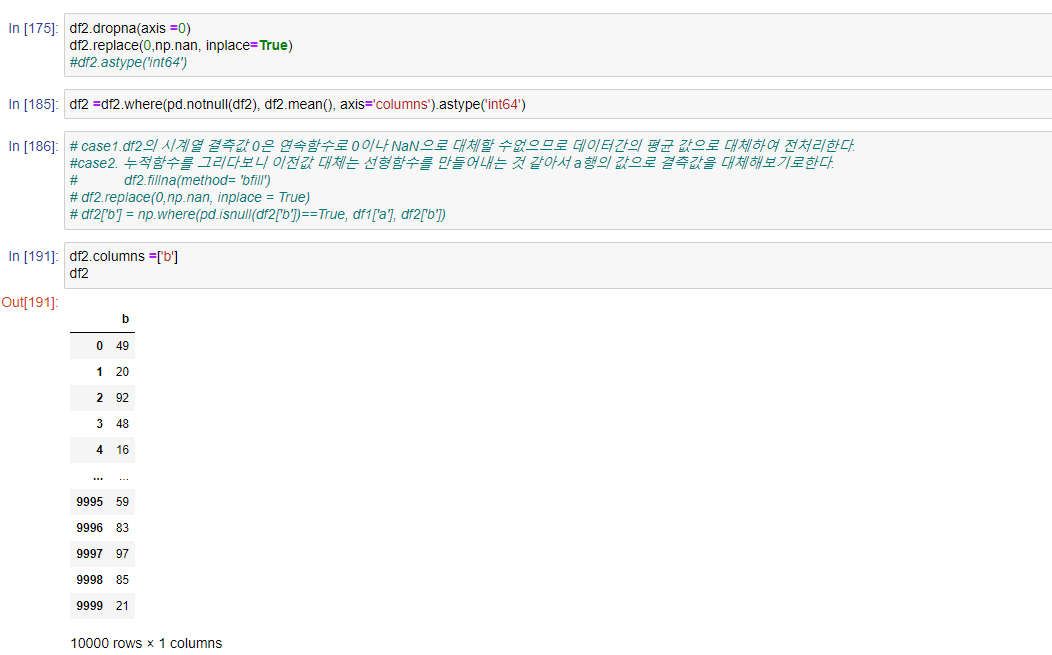
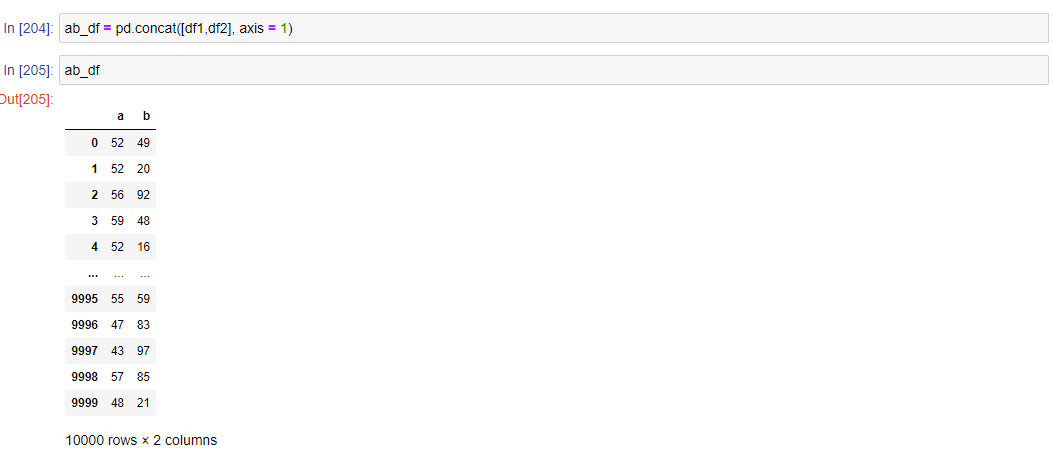
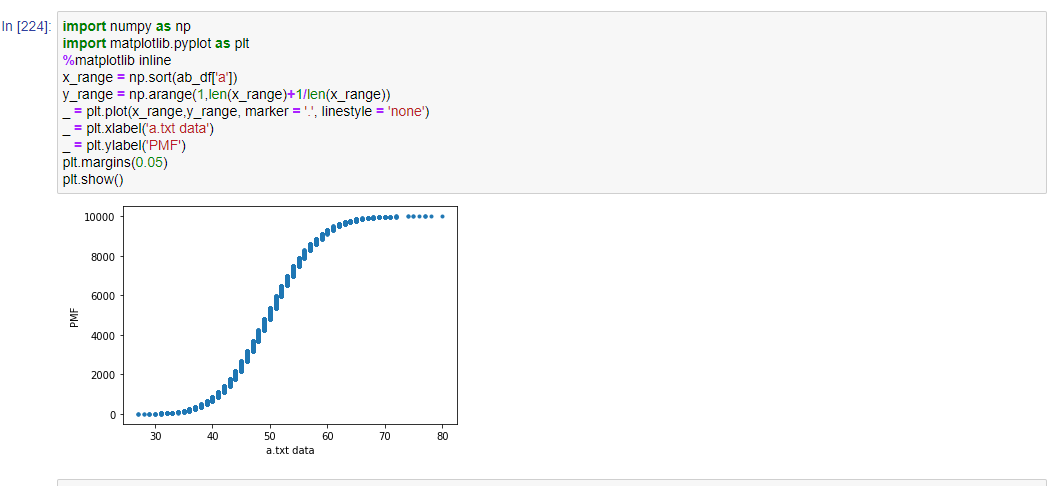
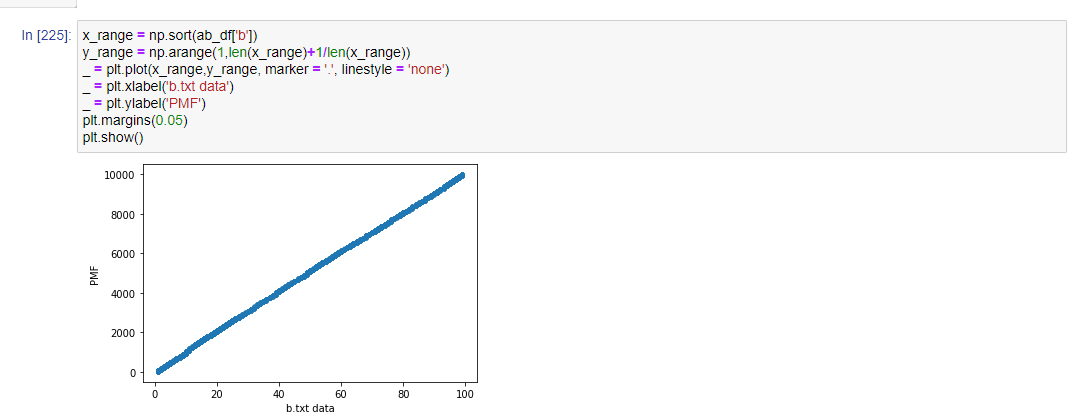
**데이터사이언스 융합학과 / 2018711118 / 오윤정/ HW5 제출**

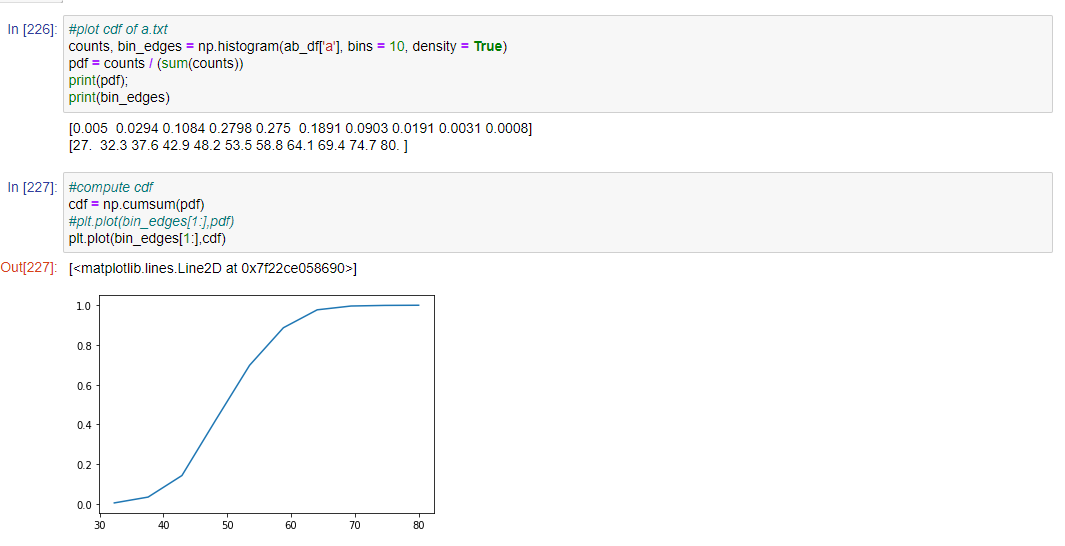
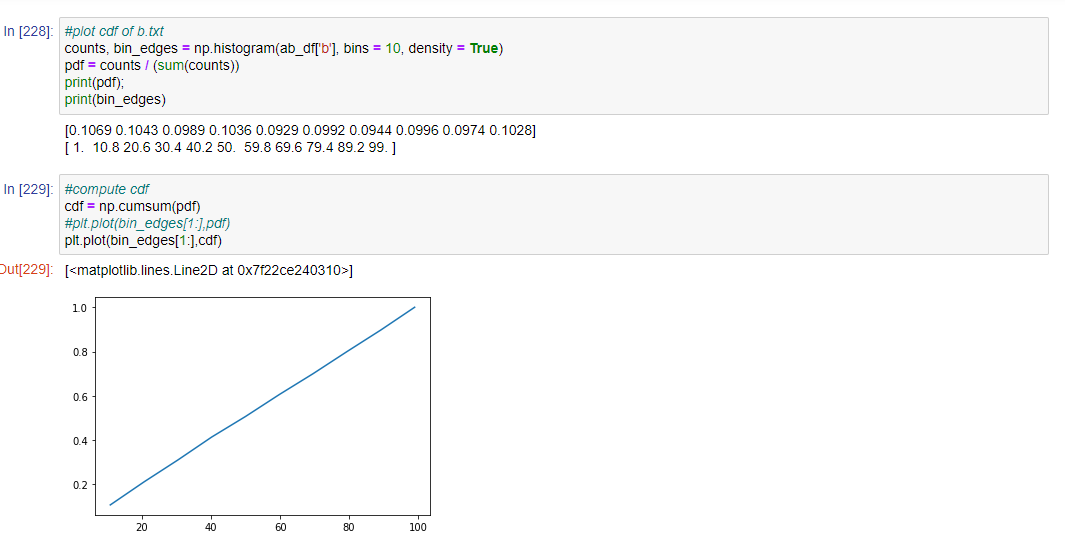
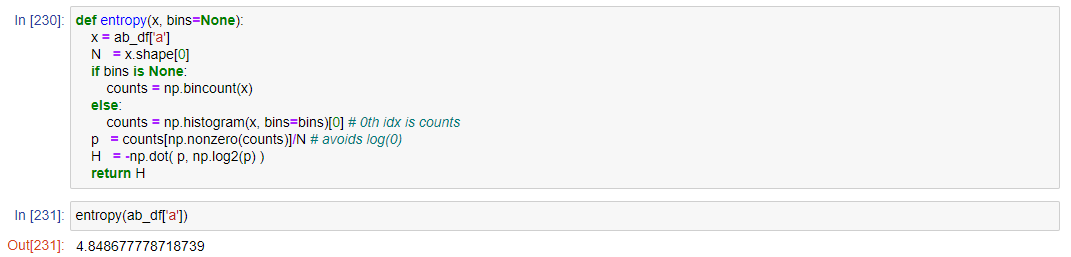
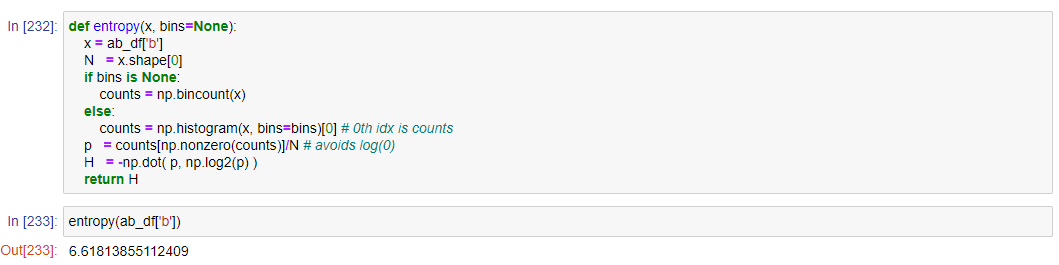
1. Check and remove any noise present in the datasets, Use the cleaned datasets for the rest of the questions.
   1. Load the datasets
   2. Import packages and NaN values dropped

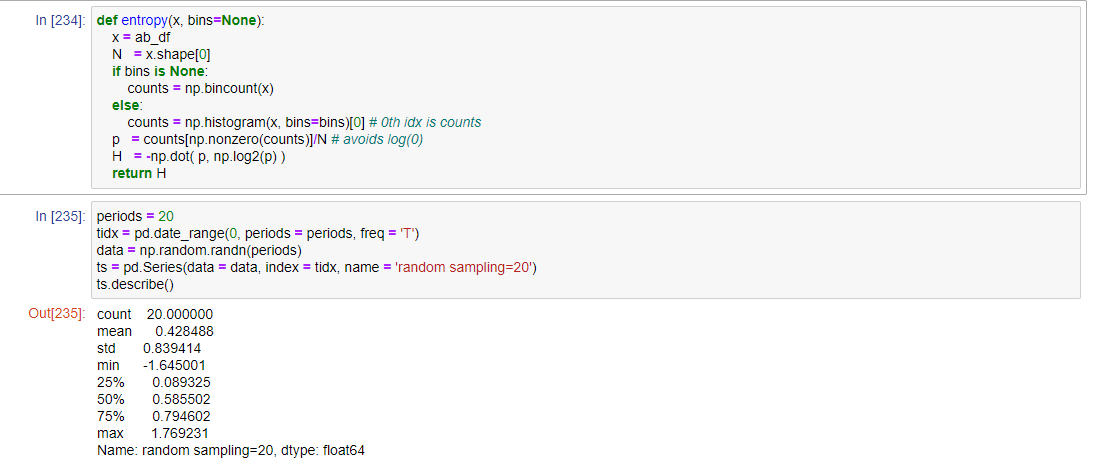


* 1. Concate the columns

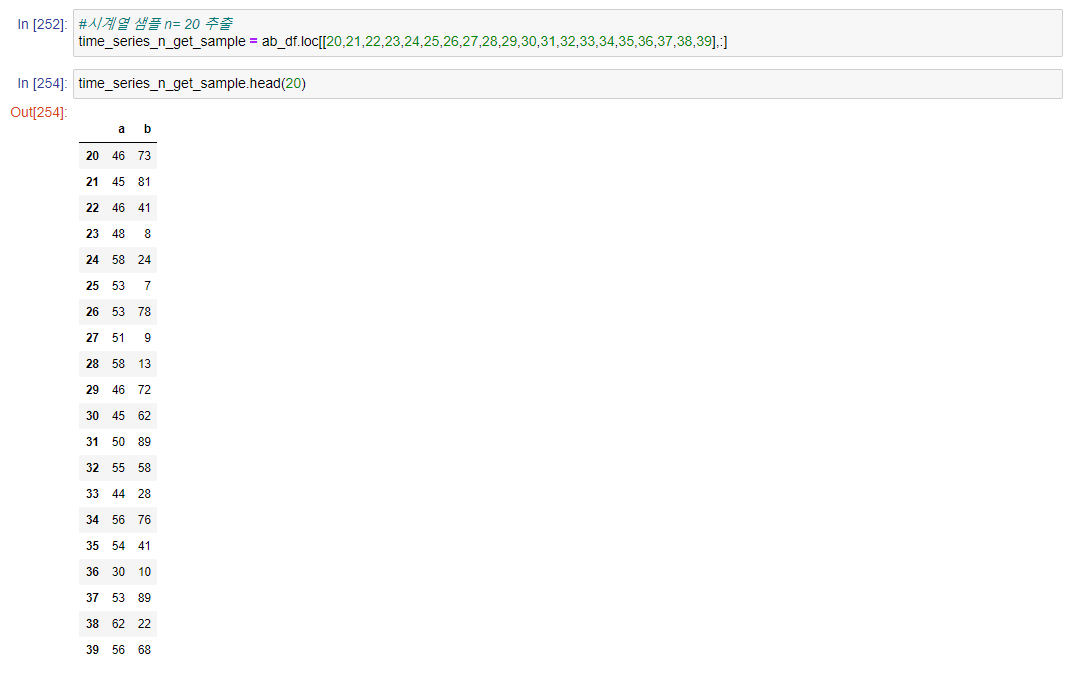
1. plot the respectie PMFs of a.txt and b.txt together for all X in [1,2,….,100]. Indicate the mean and median of each dataset in the plot(Clearly label the X-axis and the Y-axis)

## 교수님, 과제제출전에 P(0) = 2^(-100)이 P(x)가 0일 때 2^(100)으로 치환하라는 의미라고 말씀해주셧는데 pmf상의 불연속 구간에대해서 100으로 치환하라는 의미로 해석하면 되는걸까요? 그럼 pmf function에서 if구문으로 p(x) = 0값에 대한 replace를 주어야하는지? 이론적으로는 그게 가능할 것같은데 구현방법이 떠오르지 않아서 한 데까지 제출합니다.

1. Plot the respective CDFs of a.txt and b.txt together for all X in [1,2,….,100](clearly label the X-axis abd the Y-axis)
2. Compute the entropy H(P) of the random variable P(x) ~ distribution of a.txt
3. Compute the entropy H(Q) of the random variable Q(x) ~ distribution of b.txt
4. Compute H’(P,Q), for n = 20

첫 시도는 a.txt와 b.txt의 컬럼을 concat해서 묶인 하나의 ab\_df 데이터 프레임을 x값으로 같는 entropy함수를 만들어서 시계열 데이터의 period = 20값을 주어 n= 20을 추출하려했습니다.

그러나 추가적인 데이터 조작이 필요해보이고, 계산상 비효율적으로 보여 임의의 시작점에서 row를 골라 일련의 series data를 직접 추출하는 방식으로 n= 20을 추출했습니다.



1. Compute D’kl(P||Q) for n = 20. And
2. Compute D’kl(Q||P) for n= 20.

Kl divergence 함수식을 세우고, P,Q값을 입력



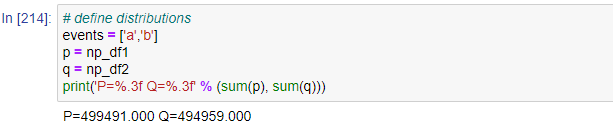
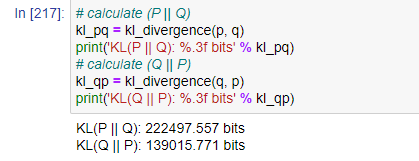
1. Compare D’kl(P||Q) and D’kl(Q||P). Are they different, same, similar? Explain
2. 위의 결과값에서도 확일 할 수 있듯이, 전혀다르다.

KL(P || Q) != KL(Q || P)

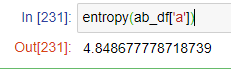
KL divergence score is not symmetrical.

As follow the definition of KL divergence, “non-symmetric difference between two probability distribution” (from wikipeida)

정보량의 차이를 KL divergence로 볼 수 있는데, p와 q의 크로스엔트로피에서 p의 엔트로피를 뺀값과 q와 p의 크로스엔트로피에서 q의 엔트로피를 뺀 값은 같을 수 없다.

1. Compute H(P) + KL(P || Q)

D’kl(P||Q) = 222497.557 bits



H(P) = 4.848…

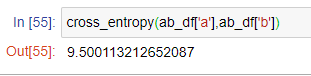
So, log(222497.557) ~=5.3473 + 4.848 = 10.19

1. Compare H(P, Q) = H(P) + KL(P || Q). are they different, same, similar? Explain.

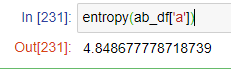
same.

H(P, Q) = H(P) + KL(P || Q)

H(P, Q)



H(P)



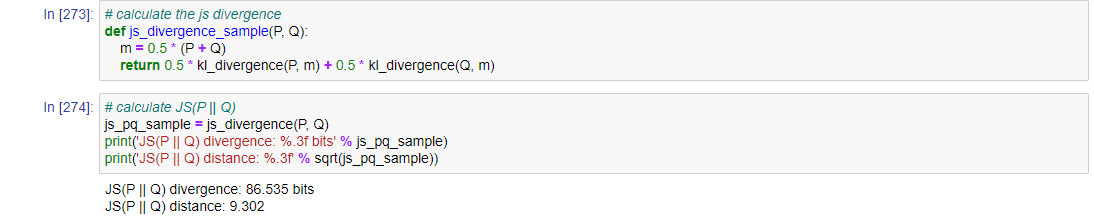
KL(P || Q)

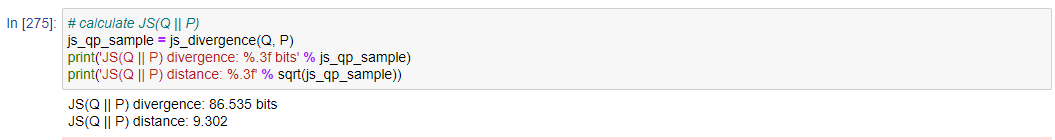


H(P, Q) = H(P) + KL(P || Q)

* 완전 동일값이 나오지 않았지만;; 근사치로 9.5 ~=10.147 값이 나왔습니다.
* 이론대로라면 same이 맞습니다. 크로스엔트로피 H(P,Q)는 엔트로피 H(P)값과 쿨백라이블러 divergence KL(P||Q)값의 합으로 이루어진 공식이기 때문입니다. 제가 계산한값으로보면 similar 로 답변해야 할 것 같습니다.

1. Compute J’(P||Q), for n = 20 where J’(P||Q\_ = 0.5 \* D’kl(P||M) + 0.5 \* D’kl(Q||M) and M = 1/2(P+Q)

Js divergence function을 만들고 값을 구한다.

1. Compute J’(Q||P), for n = 20
2. Compare J’(P||Q) and J’(Q||P). Are they different, same, similar? Explain

Same. Divergence 와 distance모두 동일하다.

Jensen shannon divergence는 수식에서도 알 수 있듯이 KL divergence의 p->q, q->p의 평균값을 내는 방식이라 symetric하고 분포간에 같은 거리를 갖는다.

따라서 J’(P||Q) and J’(Q||P)모두 동일값을 가질 수 밖에 없다.

1. Explain Why Cross Entropy is widely used as a loss function in ML learning problems.

Cross-entropy is widely used as a loss function when optimizing classification models.

Two examples that you may encounter include the logistic regression algorithm (a linear classification algorithm), and artificial neural networks that can be used for classification tasks.

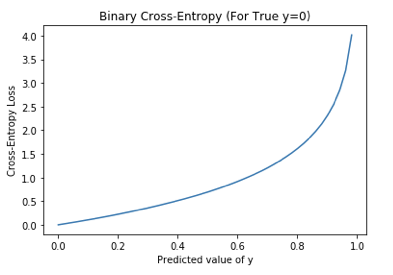
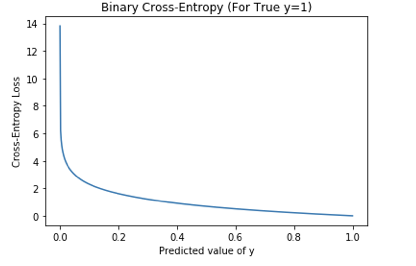
---------------------------------------------------------------------------------------------------

… using the cross-entropy error function instead of the sum-of-squares for a classification problem leads to faster training as well as improved generalization.

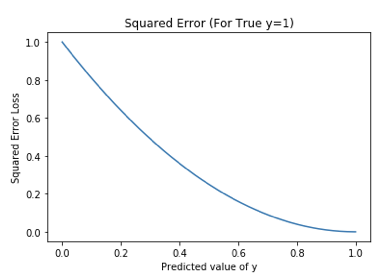
— Page 235, Pattern Recognition and Machine Learning, 2006.

사실 로스펑션의 기본적인 디폴트는 mean square error가 대표적인데, classification 문제에있어서는 대표적으로 cross entropy를 사용한다. 왜 크로스 엔트로피를 로스펑션으로 대표적으로 사용할까? 대표적인 이유로는 cross entropy는 분류 문제에 있어서 최적의 추론 확률을 return 해주는 fit function 이기 때문이다. 만일 y = 1 | 고양이 , y = 0 | ~고양이 인 분류를 할 때,

1. 크로스 엔트로피를 사용할 경우



1. Another loss function – Mean square error 사용할 경우



만일 고양이(y=1)라고 추론한경우 loss는 둘다 최소지만, panalty for guessing은 크로스엔트로피의 ‘y =0 |~고양이’ 가 훨씬 작다.

만일 내 이미지가 고양이라고 90% 확신하면 사실 1 을 예측해서 최적의 손실을 얻고, 이미지가 고양이라고 10% 확신하면 0 예측으로 가장 큰 손실을 본다.

핵심은, 크로스 엔트로피를 사용하는 것은 어떤 로스펑션도 못하는 predict true probabilities를 높인다. 그래서 category문제라면 크로스엔트로피를 사용해서 확률을 얻고 다른 로스펑션이 제공하지 못하는 최적의 추론확률을 return 받을 수 있다.

1. Explain the pros and cons of KL divergence (Dkl) as a distance metric.

Why it is useful/ not useful?

1. pros of KL divergence as a distance matric:
2. KL발산은 거리개념은 아니지만 p -> q와 q->p 가지 확률분포를 구해서 평균을 내는 방식으로 symetric as distance하게 치환해 사용할 수 있다.
3. KL divergence 는 0 이상이며, KL-divergence를 minimize하는 것이 likelihood를 maximize하는 것과 같다는 것을 활용 할 수 있다.
4. 전체를 알수없는 분포 p(x)에서 p(x) 와 q(x|θ) 사이의 KL-divergence 를 최소화시키는 θ 를 찾는 것을 통해서 학습가능한 θ의 distribution q(x|θ)를 근사 추론할 수 있다.
5. cons of KL divergence as a distance matric:

일단 쿨백라이블러 divergence는 거리개념이 아니며 non -negative의 특성을 갖고있다. 확률분포간의 거리이지만 asymetric하다.

KL(q||p)=∑q(s)logq(s)/p(s)

쿨백라이블러의 수식으로 비교해보았을때, p(s)>0과 q→0, KL의 차이가 0으로 축소되며, 이는 MLE가 시나리오에 극히 낮은 비용을 할당하고, 모델은 데이터 분포에 위치하지 않는 일부 샘플을 생성한다는 것을 의미한다. 이로 미루어보았을 때, q→0는 매우 드문확률의 경우를 의미한다. 그러면 확률은 매우크게 나타날것이다. 그런 경우 이 샘플의 분산에서의 높은 확률때문에, data training시에 부족함을 겪게된다. 이런 rare sample은 testing이나 validating시에 생성될 수있기 때문에 이점을 단점으로 보게된다.

1. why it is useful /not useful?

두 확률분포간의 거리차이를 나타내지만 p -> q와 q->p가 서로 상이하다는 점.

그러나 젠슨샤논 divergence를 사용해서 distance matric으로 쓸수 있다는 점은 유용하다.