컴퓨터비전특론 2nd week summary

[Lec03: Image classification]

Image classification

discrete labels이 가정되어 있을 때 이미지 데이터가 주어지면 해당 이미지의 label을 예측한다. 이 때 label이 n개 있다면 n x 1 probability vector가 만들어지며 vector 내에서 가장 높은 확률을 가진 label로 분류된다.

Challenges

: viewpoint variation(각도가 달라져도 알아보는지), Illumination(조명의 변화 문제), Deformation(물체 일부의 독립적인 움직임), Occlusion(가려진 영역이 있을 때의 처리), Background Clutter(지저분한 배경으로 인한 영향), Intraclass variation(동일한 카테고리 내 다른 instance)

Data-driven approach

이미지와 그에 맞는 label이 있는 dataset 수집 -> Classifier 학습 -> 새로운 이미지로 평가

Classifier: K-Nearest Neighbors

: Distance metric(ex. L1, L2 distance)을 기반으로 해당 포인트와 가까운 k개의 nearest neighborhood를 찾고, k개의 point들을 이용한 majority voting을 통해 class를 결정

=> 이 때 k와 distance metric는 hyperparameters: dataset에 따라 best value가 달라지므로 여러 경우의 수를 시도해보고 best value를 찾아야 한다. (이 때 주로 k-fold cross validation을 이용함)

\* k-fold cross validation: training set을 k개의 fold로 나누고, 돌아가면서 전체 fold 중 하나의 fold를 validation set으로, 나머지를 train set으로 사용하여 모델을 적합하고 validation set으로 모델을 평가한다. 이후 각 fold마다 계산한 error에 평균을 취하여 최종 error를 계산한다.

K-Nearest Neighbors는 매우 느리고 distance metrics가 예측하기에 충분한 정보를 담고 있지 않으며 차원이 매우 높은 문제가 발생하여 이미지 분류 문제에서는 사용되지 않는다.

Linear classification

parametric approach의 전개: Image -> -> class scores vector(# of class x 1)

이 때 가 linear classifier이며, W가 추정해야 하는 parameter(weights)

-> Weights W를 추정하기 위한 평가기준: Using loss function(parameter가 잘 추정된 정도에 대해 수치화한 값)

[Lec04: Loss Functions and Optimization]

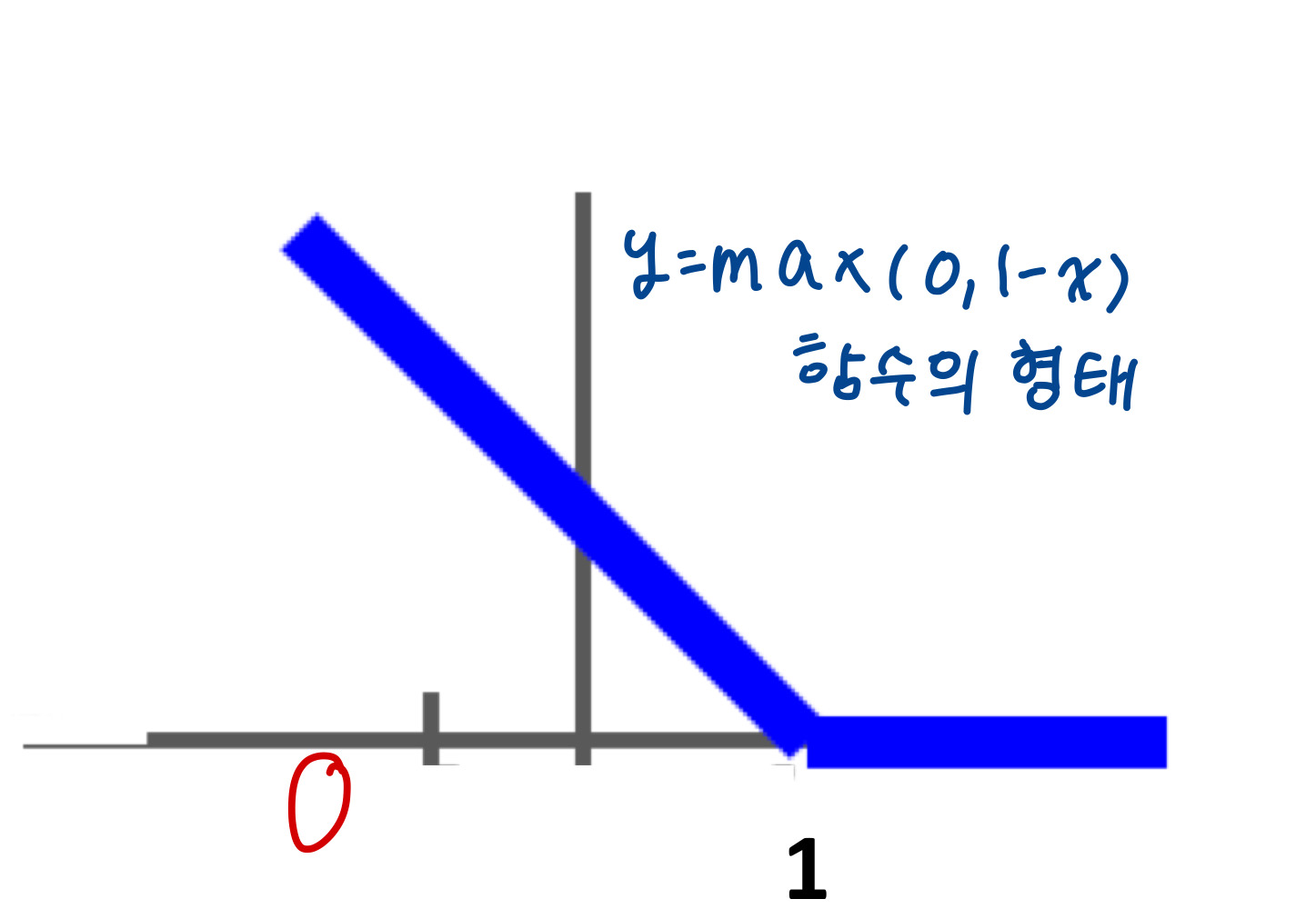
Loss function

적합한 모델을 이용하여 예측한 값이 실제 값과 얼마나 다른지에 대하여 수치화한 값.

Loss function 1) Hinge loss

- Binary hinge loss(binary SVM loss): class의 개수가 2개일 때

가 작을수록 좋은 결과라고 할 수 있다. 0보다 큰 값을 가질 수는 없으므로 가 줄기 위해서는 가 0에 가까워야 한다.의 그래프는 밑의 그림과 같으므로 일 때 가 0이 된다.



데이터의 실제 label 가 1 또는 -1이므로, positive sample은 에 해당할 것이고 negative sample은 에 해당할 것이다. 따라서 이를 좌표평면에 표시하면 positive sample의 영역과 negative sample의 영역 사이에 간격이 발생하는데, 이 간격을 margin이라고 한다. 여기서는 1 또는 -1이므로 margin은 2다.

- Hinge loss(multiclass SVM loss): class의 개수가 2개보다 많을 때

이 때 는 i번째 input의 실제 label이다. 일 때 이 되는데, 을 정리하면 이며, 이는 와 같다. 따라서 실제 label이 아닌 모든 j(j=1, 2, …, # of class)에 대해 을 만족하면 는 0이 된다. 또한 에 의하여 binary hinge loss와 마찬가지로 margin이 발생한다.

\* Class의 개수가 2개일 때 binary hinge loss로 구분되는 이유

Class의 개수가 2개일 때는 하나의 클래스에 속할 확률이 주어지면 모든 확률의 합은 1이므로 자연스럽게 나머지 클래스의 확률도 알 수 있다. 따라서 이 경우에는 최종 결과값을 scalar로 표현 가능하다. 그러나 class가 3개 이상일 경우에는 scalar로 표현이 불가하므로 두 경우의 계산방식이 달라진다.

\* Hinge loss는 정답 label을 맞추면 loss가 0으로 계산되므로, 이미 제대로 된 class가 구분된 것으로 판단하여 더 이상 계산을 하지 않는다.

Loss function 3) Log likelihood loss

는 x=1일 때 최소값을 0으로 가지므로, 가 0이 되려면 이어야 한다.

\* Softmax activation function: score를 probability의 형태로 바꿔주는 역할을 하는 함수로, 밑의 식과 같다.

위 식을 해석해보면, 기존의 score에 exponential을 취한 후 normalization을 적용하였다. Gibbs distribution에 따르면 확률은 에 비례하므로 위와 같은 전개가 가능하다.

Log likelihood loss에 softmax function을 적용하여 로 사용하는데, 이러한 형태를 softmax classifier라고 부른다.

\* Softmax Classifier vs Hinge loss

Hinge loss는 loss가 한번 0으로 계산되고 나면 더 이상 학습하지 않지만, softmax classifier는 값이 0이 되지 않아서 계속해서 학습이 가능하다. 따라서 softmax classifier가 주로 더 많이 쓰인다.

Loss function 3) Cross-entropy loss

Cross-entropy loss 또한 softmax를 적용하여 로 사용할 수 있다.

Loss function 4) Regression loss

Image를 input으로 입력했을 때 결과 값이 하나의 continuous level로 나오는 경우에 사용하는 loss로, 주로 L1 norm과 L2 norm을 사용한다.

Regularization

Train data에 과도하게 적합되는 overfitting 문제를 해결하기 위해 loss function에 regularization을 적용한다. 위 식에서

는 data loss를, 는 regularization term을 뜻한다. 이 때 는 L1, L2, elastic net 등을 사용한다. 이 때 L2 regularization의 경우 더 간소한 weight vector가 되는 방향으로 정규화가 이루어져서 weight decay라고도 불린다.

Optimization

Loss function을 최소화하는 w, b를 찾는 과정이다. 기울기를 따라 optimal value를 찾아 나가는데, 이 과정에서 negative gradient가 사용되고, loss function에 partial derivative를 적용하여 계산한다.

- Stochastic gradient descent(SGD)

: 모든 dataset에 대해 한꺼번에 gradient를 구하는 것은 계산량이 매우 커지므로, 전체 dataset에서 sampling을 통해 minibatch를 만들고 이를 이용하여 gradient를 계산한다. 이 때 sampling 과정이 추가됨으로써 gradient도 계산할 때마다 다른 값을 갖게 된다.

- Analytic gradient: linear equation

- Analytic gradient: linear equation

for a scalar x,

for a vector

- Analytic gradient: softmax activation function

softmax function in vector form ( )