Loss Functions & Optimization

$$f(x, W) = W x + b$$

구하는법? 1. Define a loss function

2. Find the parameters that minimize the loss function : optimization

Loss function

Dataset: $\{(x_i, y_i)\}_{i=1}^{N}$

image label

Loss over dataset: sum of loss over examples

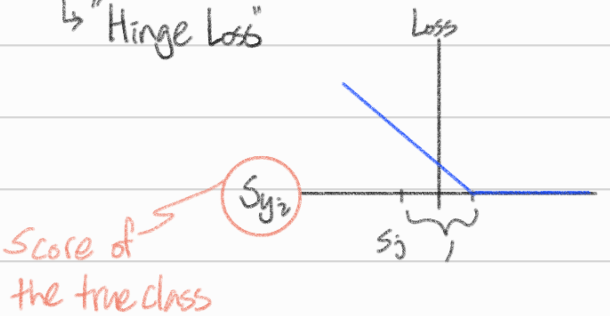
$$L = \frac{1}{N} \sum_i L_i(f(x_i, W), y_i)$$

Multiclass SVM Loss: using the shorthand for the scores vector

$s = f(x_i, W)$

$$L_i = \sum_{j \neq y_i} \begin{cases} 0 & \text{if } s_{y_i} \geq s_j + 1 \\ s_j - \underline{s_{y_i}} + 1 & \text{otherwise} \end{cases} = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1)$$

↳ "Hinge Loss"



s_j - s_{y_i} + 1

다른 class true class safety margin

이 값이 0보다 크면 그만큼이 Loss임

true class score가 다른 class score보다 월등히 커야

$$L(W) = \frac{1}{N} \sum_{i=1}^{N} L_i(f(x_i, W), y_i) + \underline{\lambda R(W)}$$

↳ Regularization for "simple" model (overfitting 방지)

L2 Regularization : Euclidean norm of weight vector W

$$R(W) = \Sigma_k \Sigma_l W_{k,l}^2$$

Softmax Classifier (Multinomial Logistic Regression)

scores = unnormalized log probabilities of the classes

$$P(Y=k \mid X=x_i) = \frac{e^{s_k}}{\Sigma_j e^{s_j}} \quad \text{where } s = f(x_i ; W)$$

$$L_i = -\log P(Y=y_i \mid X=x_i)$$ true class is close to 1

$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\Sigma_j e^{s_j}}\right)$$

ex. 

cat	3.2		$e^{3.2}$	24.5		0.13	→ $L_i = -\log(0.13)$
car	5.1	exp →	$e^{5.1}$	164.0	normalize →	0.87	$= 0.89$
frog	-1.7		$e^{-1.7}$	0.18		0.00	
	unnorm log prob.			unnorm prob.		prob.	

Now, How to find "W" that minimizes the loss

Optimization

Random Search (bad idea)

Follow the slope

↳ derivative of function

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

gradient is the vector of partial derivatives in multiple dim.

Current W	W + h	gradient dW
[0.34	[0.34 + 0.0001	[-2.5
⋮	⋮	⋮
loss 1.25347	loss 1.25322	

$$\frac{f(x+h) - f(x)}{h}$$

$$= \frac{1.25322 - 1.25347}{0.0001} = -2.5$$

위와 같은 Numerical gradient는 매우느리기 때문에 terrible idea

대신, Analytic gradient는 exact, fast해서 실제에 사용

Gradient Descent

- Full sum expensive when N is large

Stochastic Gradiet Descent (SGD 확률적 경사 하강법)

- Approximate sum using a minibatch of examples