Friday, February 5, 2021 3:04 PM

lecture 6

Vanishing Gradient Problem

$$f(x) = A\left(w_k \cdot A\left(w_{k-1} \cdot A\left(w_{k-2} \cdot A(\cdots A(w_1 x)\cdots\right)\right)\right)\right)$$

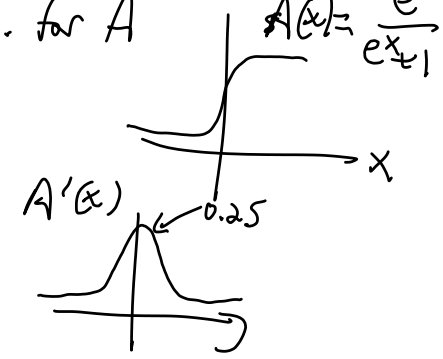
$$\frac{\partial f}{\partial w_k} = G_k x_k$$

$$x_k = A(w_{k-1} \cdot x_{k-1})$$

$$\boxed{G_{k-1} = G_k w_k A'(w_{k-1} x_{k-1})}$$

In early days of NN's, common to use sigmoid fn. for A

$$A(x) = \frac{e^x}{e^x + 1}$$



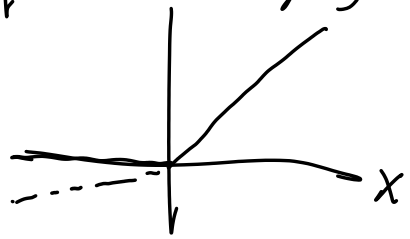
weights getting stuck far from optimum ←

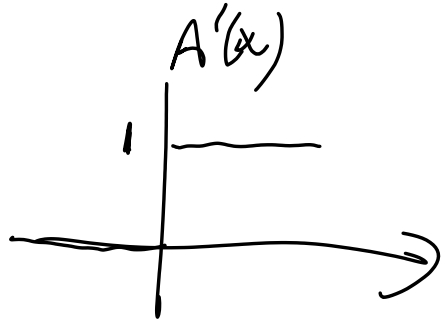
decreasing or suppressed gradients esp. for initial layers. as network gets deeper. ←

Friday, February 5, 2021 3:32 PM

Sol'n: use a better activation fn

e.g. ReLU: $\max(0, x)$





enabled much deeper networks

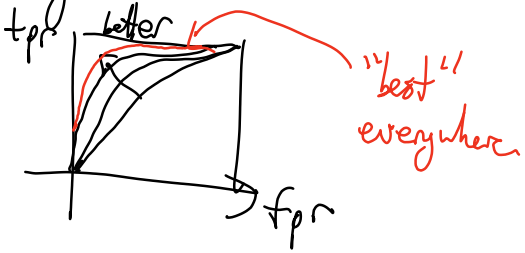
Friday, February 5, 2021 3:37 PM

Back to binary classification

Key result in binary classification: Neyman–Pearson Lemma : optimal binary classifier is the Likelihood Ratio

$$R(x) = \frac{P(x|S)}{P(x|B)}$$

uniformly best ROC curve.



at fixed tpr $R(x)$ achieves lowest possible fpr for every tpr.

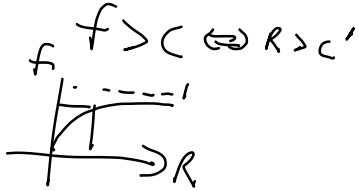
Proof: consider BCE loss

$$L = \sum_{i\in S} \log f(x) + \sum_{i\in B} \log(1-f(x))$$

$$= \int dx \left[P(x|S) \log f(x) + P(x|B) \log(1-f(x)) \right]$$

$f \to f + \delta f$, set 1st der. to zero

$$0 = \frac{P(x|S)}{f} - \frac{P(x|B)}{1-f} \longrightarrow \boxed{f = \frac{P(x|S)}{P(x|S)+P(x|B)}} = \frac{R(x)}{R(x)+1}$$



$\left(f(x) = P(S|x) \,?\right)$

Bayes thm: $P(S|x)P(x) = P(x|S)P(S)$

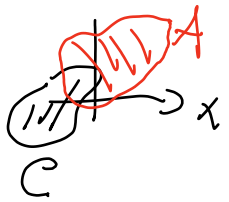
Friday, February 5, 2021 3:49 PM

f is monotonic in R $\Rightarrow$ equivalent classifiers ✓

LR minimizes BCE loss ✓.

Proof #2: (directly showing that R gives best possible ROC curve)

Consider any other classifier $f(x)$. $f(x) > c$ $\rightarrow$ defines a region C in x space



$\downarrow$

$$tpr = \varepsilon_S(f) = \int_C P(x|S)$$

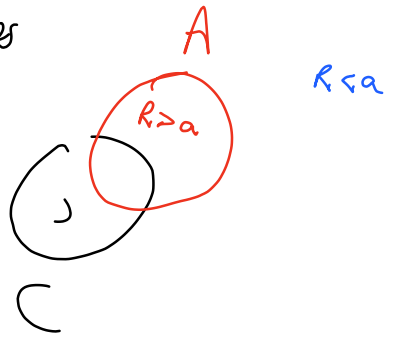
Now consider a cut on $R(x) > a$ w/ same tpr. This defines another region A

$$\varepsilon_B(f) = \varepsilon_B(R) = \int_A P(x|B).$$

Friday, February 5, 2021 3:56 PM

We want to compare $\varepsilon_S(f)$ vs $\varepsilon_B(R)$. (background efficiencies or fprs)

$$\int_C P(x|B) \qquad \int_A P(x|B).$$



$$\varepsilon_B(R) = \int_A P(x|B) = \int_{A\cap C} P(x|B) + \int_{A\cap C^\perp} P(x|B)$$

$$= \int_C P(x|B) - \int_{A^\perp\cap C} P(x|B) + \int_{A\cap C^\perp} P(x|B)$$

$$\left(R(x) = \frac{P(x|S)}{P(x|B)}\right)$$

$$= \varepsilon_B(f) - \int_{A^\perp\cap C} \frac{1}{R(x)} P(x|S) + \int_{A\cap C^\perp} \frac{1}{R(x)} P(x|S)$$

$$< \varepsilon_B(f) + \frac{1}{a}\left[\int_{A\cap C^\perp} P(x|S) - \int_{A^\perp\cap C} P(x|S)\right] + \frac{1}{a}\left[\int_{A\cap C} P(x|S) - \int_{A\cap C} P(x|S)\right]$$

$$= \varepsilon_B(f) + \frac{1}{a}\int_A P(x|S) - \frac{1}{a}\int_C P(x|S)$$

ε !

$$= \varepsilon_B(f) \checkmark$$

Friday, February 5, 2021 4:02 PM

NP lemma is extremely useful!

- optimal classifier exists, given by LR. → provide strategy for hypothesis testing, e.g. search for new physics
  construct likelihood for signal & background.

- modern ML: "likelihood ratio trick"

Assume: NN can learn the optimal classifier → $\frac{P(x|S)}{P(x|B)}$

"likelihood free inference" ← gives access to likelihood ratio w/out knowing indiv. likelihoods!

Applications
- generation: GANs
- anomaly detection: "weak supervision", CWoLa
- phase space reweighting e.g. simulation vs data
⋮

Friday, February 5, 2021 3:56 PM

Multiclass Classification

- Important & straightforward generalization of binary classification

- Objective: learn class probabilities $p(C_k|x)$ $k=1,\dots N_{class}$

$$\sum_k p(C_k|x) = 1.$$

Ex: MNIST classify 0–9 from each other.

- loss fn? MLE binary: $\sum_{i\in S} \log p(S|x) + \sum_{i\in B} \log p(B|x)$

multiclass:

$$L = \sum_k \sum_{i\in C_k} \log p(C_k|x)$$

"categorical cross entropy"

Friday, February 5, 2021 4:27 PM

Think of $p(C_k|x) \xrightarrow{\text{vector}} \begin{pmatrix} p(C_1|x) \\ \vdots \\ p(C_N|x) \end{pmatrix} = \vec{p}(x)$

introduce truth labels

$k=1$ $\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\leftarrow y$ for $k=1$, $k=2$ $\begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\leftarrow y$ for $k=2$, ... $k=N$ $\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$ $\leftarrow y$ for $k=N$

"one-hot encoding"

$$L = \sum_{i \in \text{data}} \vec{y}_i \cdot \log \vec{p}(x_i)$$

log acts elementwise over vector $\vec{p}$.

truth label for data pt. i

Friday, February 5, 2021 4:33 PM

Metrics:

- Accuracy: fraction of $\operatorname{argmax}\ \vec{p}(x) = \text{true class}$ "top-1 accuracy"

"top-n accuracy" → correct label in n highest $\vec{p}$'s.

- one-vs-rest binary metrics
- confusion matrix