CMPUT 367

Roderick Lan

Contents

Lecture 5 - Jan 25

1 Assignment

softmax: $y_k = f\dfrac{\exp\{z_k\}}{\sum_{k=1}^{K}\exp\{z_k\}}$

$x \in \mathbb{R}^d$
$y \in \mathbb{R}^k$
$z \in \mathbb{R}^k$
$W \in \mathbb{R}^{k \times d}$
$y = \text{softmax}(z) = \text{softmax}(Wx + b)$

$$\frac{z}{k} = \frac{W}{k \times d}\cdot$$

W, rows are vectors

2 Lecture

Log Reg:
$z = w^\top x + b$
$y = \sigma(z) = \dfrac{1}{1+e^{-(z)}}$
Classification (Softmax):
$z_k = w_k^\top x + b_k$
$y_k = \dfrac{\exp(z_k)}{\sum_{k'}\exp(z_{k'})}$
If $k = 2$, softmax/classification and log reg are equivalent
(proof of equality in thursday, jan 25 lecture notes)
sigmoid function comes from softmax; divide numerator
bias term b is learned

2.1 How To Train Model

Minimize loss function

$$\underset{w,b}{\text{minimize}} J(w,b)$$

In logistic regression, use cross entropy loss:

$$J = \sum_{m=1}^{M}\left[-t^{(m)}\ln y^{(m)} - (1-t^{(m)})\ln(1-y^{(m)})\right]$$

In softmax (multiclassification), use cross entropy loss for multiple classes:

$$J = \sum_{m=1}^{M}\left[-\sum_{k=1}^{K} t_k^{(m)}\ln y_k^{(m)}\right]$$

CE loss is also the same for logistic regression and 2-way classification
No closed form solution for log reg due to exp (cant group/separate unknown from known var)

2.1.1 Logistic Regression

$$\frac{\partial J}{\partial w_i} = \frac{\partial}{\partial w_i} \sum_{m=1}^{M} \left[-t^{(m)} \ln y^{(m)} - (1 - t^{(m)}) \ln(1 - y^{(m)}) \right]$$

$$\frac{\partial J^{(m)}}{\partial w_i} = \frac{\partial}{\partial w_i} \left[-t \ln y - (1 - t) \ln(1 - y) \right]$$

m omitted; need to expand y (contains w)

$$\begin{aligned}
&= \frac{\partial}{\partial w_i} \left[-t \ln \frac{1}{1 + e^{-(w^\top x + b)}} - (1 - t) \ln(1 - \frac{1}{1 + e^{-(w^\top x + b)}}) \right] \\
&= \frac{\partial}{\partial w_i} \left[t \ln(1 + e^{-(w^\top x + b)}) - (1 - t)(\ln e^{-z} - \ln(1 + e^{-z})) \right] \\
&= \frac{\partial}{\partial w_i} \left[t \ln(1 + e^{-(w^\top x + b)}) - (1 - t)(-z - \ln(1 + e^{-z})) \right] \\
&= \frac{\partial}{\partial w_i} \left[t \ln(1 + e^{-z}) - (1 - t)(-z - \ln(1 + e^{-z})) \right] \\
&= \frac{\partial}{\partial w_i} \left[-(1 - t)(-z) - 1 \cdot (-\ln(1 + e^{-z})) \right] \\
&= \frac{\partial}{\partial w_i} \left[(1 - t)z + \ln(1 + e^{-z}) \right] \\
&= \frac{\partial}{\partial w_i} \left[(1 - t)(w^\top x + b) + \ln(1 + e^{-(w^\top x + b)}) \right] \\
&= (1 - t) \cdot x_i + \frac{1}{1 + e^{-z}} e^{-z}(-x_i) \\
&= (1 - t)x_i - \frac{e^{-z}}{1 + e^{-z}} x_i \\
&= \left(1 - \frac{e^{-z}}{1 + e^{-z}} - t \right) x_i = \left(\frac{1}{1 + e^{-x}} - t \right) x_i = (y - t)\, x_i
\end{aligned}$$

y - prediction
t - target
x_i - amplified by feature

2.1.2 Linear Regression

$$J^{(m)} = \frac{1}{2}(y - t)^2 \quad \text{where } y = w^\top x + b$$

$$\frac{\partial J^{(m)}}{\partial w_i} = (y - t)x_i$$

Same as log reg. Not a coincidence.
Generalized Linear Models (GLIM)

2.1.3 Softmax/Multiclassification

Still a linear model, squashing function not linear
$z = Wx + b$
$y = \text{softmax}(z)$
$W \in \mathbb{R}^d$

$$\frac{\partial J}{\partial w_{ij}} = (y_i - t_i)x_j$$

i - category; j - feature
y goes with category, x goes with feature

2.2 Convexity

Definition 2.2.1: Convex Set

A set S is a convex set iff $\forall x, y \in \mathcal{S}$, $\forall \lambda \in (0, 1)$

$$\lambda x + (1 - \lambda)y \in \mathcal{S}$$

Every point is trying to "stretch out", no point folds into interior.
Draw a line b/w points $\rightarrow$ all points on line should be part of set.
Disconnected sets are never convex sets.

Definition 2.2.2: Convex Function

A real valued function f is convex iff

1. domain of f is a convex set
2. $\forall x, y \in$ domain f, $\forall \lambda \in (0, 1)$

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(\lambda x + (1 - \lambda)y)$$

2.3 Optimization

$$w \leftarrow w + \alpha \nabla f$$