Week 2: Tutorial Hand Note

Unit Vector $\hat{\mathbf{v}}$/单位向量

定义:

$$\vec{v} \text{ is } \hat{v} \text{ iff. } \left\|\vec{v}\right\| = 1$$

即长度为 1 的向量为单位向量。

Norm 2

$\|\vec{v}\|$ 即向量 v 的 2-范数（Norm 2，即欧几里得范数，或者说是几何距离），其定义为

$$\left\|\vec{v}\right\| = \sqrt{\sum_{i=1}^{d} v_i^2}$$

巧思：这个也同样是 L2 Loss 的表达形式

Normalisation

对于任意向量（除 0 以外），其总能获得与其方向（Direction）相同的单位向量。其公式为：

$$\|\hat{v}\| = \frac{\vec{v}}{\|v\|}$$

我们称上述公式为 Normalisation。

Taylor Polynomials/泰勒级数

单变量公式： $\mathbb{R} \mapsto \mathbb{R}$

对于函数 $E(x)$ 于 w_0 处展开，获得 w 处的近似值，其公式被定义为：

$$\mathrm{T}(w) = \sum_{k=0}^{n} \left[\frac{E^{(k)}(w_0)}{k!} (w - w_0)^k \right]$$

多变量公式：$\mathbb{R}^d \mapsto \mathbb{R}$

$$\mathrm{T}(\vec{w}) = \sum_{k=0}^{n} \left[\frac{E_{\vec{w}}^{(k)}(\overrightarrow{w_0})}{k!} (\vec{w} - \overrightarrow{w_0})^k \right]$$

其中$E_{\vec{w}}^{(k)}$,(k)表示 k 阶导数，而$\vec{w}$ 表示是对函数 $E(w)$ 的全导数（total derivative）。

$$\begin{aligned} E_{\vec{w}}^{(0)}(\overrightarrow{w_0}) &= E(\overrightarrow{w_0}) \\ E_{\vec{w}}^{(1)}(\overrightarrow{w_0}) &= \nabla E(\overrightarrow{w_0}) \\ E_{\vec{w}}^{(2)}(\overrightarrow{w_0}) &= \mathbf{H}(\overrightarrow{w_0}) \end{aligned}$$

$$\begin{aligned} &(\vec{w} - \overrightarrow{w_0})^0 = 1 && 1 \\ &(\vec{w} - \overrightarrow{w_0})^1 = (\vec{w} - \overrightarrow{w_0}) && d \times 1 \\ &(\vec{w} - \overrightarrow{w_0})^2 = (\vec{w} - \overrightarrow{w_0}) \bullet (\vec{w} - \overrightarrow{w_0})^{\mathrm{T}} && (d \times 1) \times (1 \times d) = d \times d \end{aligned}$$

$$\begin{aligned} E_{\vec{w}}^{(k)}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0})^k &= ? \\ E_{\vec{w}}^{(0)}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0})^0 &= E(\overrightarrow{w_0}) \\ E_{\vec{w}}^{(1)}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0})^1 &= \nabla E(\overrightarrow{w_0})^{\mathrm{T}}(\vec{w} - \overrightarrow{w_0}) \\ E_{\vec{w}}^{(2)}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0})^2 &= (\vec{w} - \overrightarrow{w_0})^{\mathrm{T}} \mathbf{H}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0}) \end{aligned}$$

How to prove that GD really takes a step in the direction of the steepest descent?

如何证明 "梯度下降法 "真的向最陡下降的方向迈出了一步?

可以将 GD 的公式记为

$$\vec{w} = \vec{w} - \eta \nabla E(\overrightarrow{w_0})$$

假设其的公式为

$$\vec{w} = \overrightarrow{w_0} + \eta \hat{v}$$

其中 η 是学习率，也同样是学习的步长（每一次学习，学习多长）

而 $\vec{v}$ 则是学习的方向（向哪里学习？）

→ 因此我们需要找到一个 $\boldsymbol{\vec{v}}$ 使 $\boldsymbol{E(w) - E(w_0)}$ 尽可能的负（Negative）

解释： 因为是学习，所以我们期望学习后的 $E(w)$ 尽可能小（当然，需要小于$E(w_0)$）

所以我们期望 $E(w) - E(w_0) \leq 0$

而如果我们学习方向足够正确，那么一定会让$E(w)$足够小，也就意味着

$$(E(w) \downarrow - E(w_0)) \downarrow$$

也就是期望其尽可能的负。

→ 因此我们需要证明 $\boldsymbol{-\nabla E(\overrightarrow{w_0})}$ 是尽可能的负

假设： 学习率 η 足够小

我们对 $\mathrm{E}(\vec{\mathrm{w}})$ 进行 1 阶泰勒展开

$$\mathrm{E}(\vec{\mathrm{w}}) = \frac{E_{\vec{w}}^{(0)}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0})^0}{0!} + \frac{E_{\vec{w}}^{(1)}(\overrightarrow{w_0})(\vec{w} - \overrightarrow{w_0})^1}{1!} + \cdots (\text{ignored})$$

$$\mathrm{E}(\vec{w}) = E(\overrightarrow{w_0}) + \nabla E(\overrightarrow{w_0})^{\mathrm{T}}(\vec{w} - \overrightarrow{w_0})$$
$$\mathrm{E}(\vec{w}) - E(\overrightarrow{w_0}) = \nabla E(\overrightarrow{w_0})^{\mathrm{T}}(\vec{w} - \overrightarrow{w_0})$$

代入 $\vec{w} = \overrightarrow{w_0} + \eta \hat{v}$

$$\begin{aligned} E(\overrightarrow{w_0} - \eta \vec{v}) - E(\overrightarrow{w_0}) &= \nabla E(\overrightarrow{w_0})^{\mathrm{T}}(\overrightarrow{w_0} + \eta \hat{v} - \overrightarrow{w_0}) \\ &= \eta \nabla E(\overrightarrow{w_0})^{\mathrm{T}} \overrightarrow{w_0} \end{aligned}$$

End of the Tutorial, TBC.