Week 2: Tutorial Hand Note

Unit Vector $\hat{v}$/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：$\mathbb{R}^d \mapsto \mathbb{R}$

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

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）。

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

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

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

其中：

$$\nabla E(\mathbf{w}) = \begin{pmatrix} \dfrac{\partial E(\mathbf{w})}{\partial \mathrm{w}_1} \\ \dfrac{\partial E(\mathbf{w})}{\partial \mathrm{w}_2} \\ \vdots \\ \dfrac{\partial E(\mathbf{w})}{\partial \mathrm{w}_{d-1}} \\ \dfrac{\partial E(\mathbf{w})}{\partial \mathrm{w}_d} \end{pmatrix}$$

$$\begin{aligned}
&\mathbf{H}\big(E(\mathbf{w})\big) \\
&= \mathbf{H}_E \\
&= \begin{bmatrix}
\dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_1^2} & \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_1 \partial \mathrm{w}_2} & \cdots & \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_1 \partial \mathrm{w}_d} \\
\dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_2 \partial \mathrm{w}_1} & \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_2^2} & \cdots & \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_2 \partial \mathrm{w}_d} \\
\vdots & \vdots & \ddots & \vdots \\
\dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_d \partial \mathrm{w}_1} & \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_d \partial \mathrm{w}_2} & \cdots & \dfrac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_d^2}
\end{bmatrix}
\end{aligned}$$

$$(\mathbf{H}_E)_{i,j} = \frac{\partial^2 E(\mathbf{w})}{\partial \mathrm{w}_i \partial \mathrm{w}_j}$$

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 $\vec{v}$ 使 $E(w) - E(w_0)$ 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

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

也就是期望其尽可能的负。

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

假设： 学习率 η 足够小

假设原因： 我们需要使用泰勒级数进行展开

我们对 E($\vec{w}$) 进行 1 阶泰勒展开

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

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

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

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

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

$$= \eta \nabla E(\vec{w_0})^{\mathrm{T}} \hat{v}$$

End of the Tutorial (Thursday, Week 2), TBC.

Following Content is Week2, Friday, the Normal Lecture.

整理上式，我们可以得到：

$$E(\vec{w_0} - \eta\vec{v}) - E(\vec{w_0}) = \eta \nabla E(\vec{w_0})^{\mathrm{T}} \hat{v}$$

并且我们期望$E(\vec{w_0} - \eta\vec{v}) - E(\vec{w_0})$尽可能的负。

→ 因此我们期望 $\boldsymbol{\eta \nabla E(\vec{w_0})^{\mathrm{T}} \vec{w_0}}$ 尽可能的负

→ 不考虑常数项 $\boldsymbol{\eta}$，因此我们期望 $\boldsymbol{\nabla E(\vec{w_0})^{\mathrm{T}} \vec{w_0}}$ 尽可能的负

而对于 $\nabla E(\vec{w_0})^{\mathrm{T}} \vec{w_0}$，我们可以将其看作向量 $\nabla E(\vec{w_0})^{\mathrm{T}}$ 与向量 $\vec{w_0}$ 的点乘（Dot Product）。因此我们可以将上式写成

$$\begin{aligned} \nabla E(\vec{w_0})^{\mathrm{T}} \hat{v} &= \left\| \nabla E(\vec{w_0})^{\mathrm{T}} \right\| \left\| \vec{v} \right\| \cos\theta \\ &= \left\| \nabla E(\vec{w_0}) \right\| \cos\theta \end{aligned}$$

其中 θ 表示向量 $\hat{v}$ 与梯度 $\nabla E(\vec{w_0})$ 的夹角。

因此我们可以改变上式得到：

$$\cos\theta = \frac{\nabla E(\vec{w_0})^{\mathrm{T}} \hat{v}}{\|\nabla E(\vec{w_0})\|} \in [-1,1]$$

我们知道 $\cos\theta \in [-1, 1]$，因此

$$(\theta = 180°) \quad -1 \le \cos\theta = \frac{\nabla E(\vec{w_0})^{\mathrm{T}} \hat{v}}{\|\nabla E(\vec{w_0})\|} \le 1 \quad (\theta = 0°)$$

$$(\theta = 180°) \quad -\|\nabla E(\vec{w_0})\| \le \nabla E(\vec{w_0})^{\mathrm{T}} \hat{v} \le \|\nabla E(\vec{w_0})\| \quad (\theta = 0°)$$

因此

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

因为我们期望 $\nabla E(\vec{w_0})^{\mathrm{T}} \vec{w_0}$ 尽可能的负，因此我们使其取最小值 $-\|\nabla E(\vec{w_0})\|$

而此时，表示方向的单位向量 $\hat{v}$ 与梯度 $\nabla E(\vec{w_0})$ 的夹角 θ 为 180°。

因此我们学习时候的 $\hat{v}$ 的方向与 $-\nabla E(\vec{w_0})$ 一致。

这证明了"梯度下降法 "真的向最陡下降的方向迈出了一步。