复习：Linear regression

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})^2$$

$$= \frac{1}{2n}\|y - Xw\|^2$$

$$\nabla L(w) = \frac{1}{n}(-X^T y + X^T X w)$$

$$\left(= \frac{1}{n}\sum_{i=1}^{n}\nabla_w (y^{(i)} - w^T x^{(i)})^2\right)$$

2-1 ~ 2-2：为什么选择MSE，为什么MSE得出的parameters为 best linear unbiased estimator.

2-1 Likelihood

注意到：这里的 loss function $L(w)$ 源于 MSE (Mean squared error)

MSE 的 loss function 可用于 linear regression 的原因是我们假设了观测中包括服从 Gaussian distribution 的 error ε.

即：实际的 $y = \vec{w}^T\vec{x} + b + \varepsilon$

其中 $\varepsilon \sim N(0, \sigma^2)$，即服从 Gaussian distribution.

也就是 $p(\varepsilon) = \frac{1}{\sqrt{2\pi\sigma_{(\varepsilon)}^2}}\exp\left(-\frac{1}{2\sigma_{(\varepsilon)}^2}(\varepsilon - \mu_{(\varepsilon)}^2)\right)$

在这一假设下，MSE 得出的参数是 best linear unbiased estimator (BLUE).

此时 model 对于一个给定的 data point $\vec{x}$，观测到某个结果 y 的 likelihood 为：

$$p(y|\vec{x}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y - \vec{w}^T\vec{x} - b)^2\right)$$

likelihood，似然，即在给定一个 x 下观测到 y 的概率. 不同的 y 有不同的概率，然而当我们再固定住 y，likelihood 就变为：对于观测到的一组 $\vec{x}$ 和它的观测结果 y，在我们的参数 $\vec{w}, b$ 下，模型也会得到这个 y 的概率.

实际 $p(y|\vec{x})$ 可写作 $L(\vec{w}, b|\vec{x})$，这才是这个函数的目的.

即其实：$p(y|\vec{x})$ 是一个在某组数据 $(\vec{x}, y)$ 给定下，model 能正确地从 $\vec{x}$ 推出 y 的概率 p 关于参数 $(\vec{w}, b)$ 的 function. Input 为一组参数 $(\vec{w}, b)$，output 为概率 p.

所以很显然，我们需要 model 对任何输入 x 尽量更可能给出正确的 output y，因而我们需要找出使所有 $p(y|\vec{x})$ 值尽可能最大的 $\vec{w}, b$.

很显然，几乎不可能有一组 $\vec{w}, b$ 使对于每组 $(\vec{x}, y)$ 都有最大的 likelihood $p(y|\vec{x})$ 但是我们可以找出对于整个 dataset (assuming independence among samples) 综合 likelihood 最大的 parameter $(\vec{w}, b)$

$$p(\vec{y}|X) = \prod_{i=1}^{n} p(y^{(i)}|\vec{x}^{(i)}) \quad \text{(*)}$$

maximum likelihood principle：找到最佳 parameter $(\vec{w}, b)$ which maximize $p(\vec{y}|X)$.

这一过程叫 maximum likelihood estimation (MLE, 最大似然估计)

2-2 如何寻找 MLE 的最佳 parameter $(\vec{w}, b)$

我们可以 ① take log (这是一个 monotonic transformation)

② take minus sign (不改变数据顺序. 这样求 max 就转为求 min.)

这样一来求 MLE 就等价于 minimizing the negative log likelihood

这种方法最大的好处是用 log 把 $\prod$ 转为了 $\sum$，求 max 转为求 min (much easier)

$$-\ln\left(p(\vec{y}|X)\right) = -\ln\left(\prod_{i=1}^{n} p(y^{(i)}|\vec{x}^{(i)})\right)$$

$$= \sum_{i=1}^{n} -\ln p(y^{(i)}|\vec{x}^{(i)})$$

$$= \sum_{i=1}^{n} -\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - w^T\vec{x}^{(i)} - b)^2\right)\right)$$

$$= \sum_{i=1}^{n} \underbrace{\frac{1}{2}\log(2\pi\sigma^2)}_{\text{const}} + \underbrace{\frac{1}{2\sigma^2}}_{\text{const}}(y^{(i)} - w^T\vec{x}^{(i)} - b)^2$$

因而，MLE is equivalent to minimize MSE (*)

这就是为什么 MSE 是 best linear unbiased estimator.

以上就是 linear regression 的所有 theoratical part

整理：① 我们需要一个使 $P(y|X)$ 最大的 loss function

⇒ ② 使用 MSE 作为 loss function

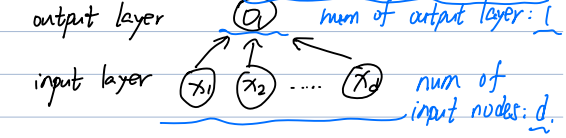
⇒ ③ 要使 MSE 最小，可以有 analytic sol，

也可使用 gradient descend.

2-3. Neural network.

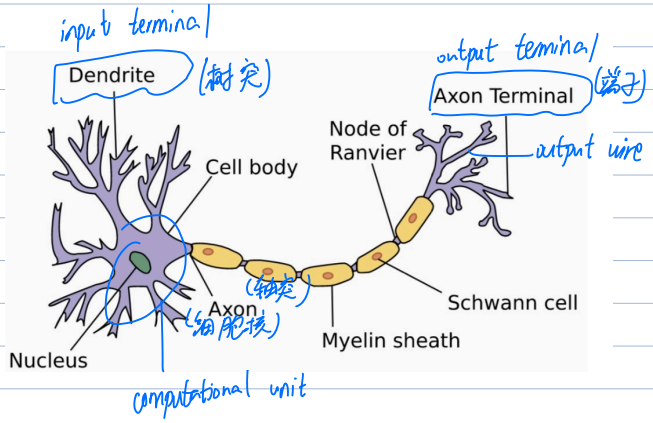
Linear Regression 是一个 single layer neural network. (layer num = 1)

(通常说层数不包括 input layer)



此处每个 input 只和一个 output 相连. (全连接层)

这种 output layer 叫做 fully-connected layer 或者 dense layer (稠密层)



① 上一个 unit 的 axon terminal 将 通过 synapse (突触) 传入这个 unit 的 dendrite 作为 input.

② synapse 承载 weights，经过 nucleus (computational unit) 进行运算. + weights 作为 activation, − weights 作为 inhibition.

③ 传输结果 y 到 axon，通常 axon 会做一些 nonlinear 的处理 $\sigma(y)$

④ 结果传入这个 unit 的 axon terminal，准备传入下个 unit.