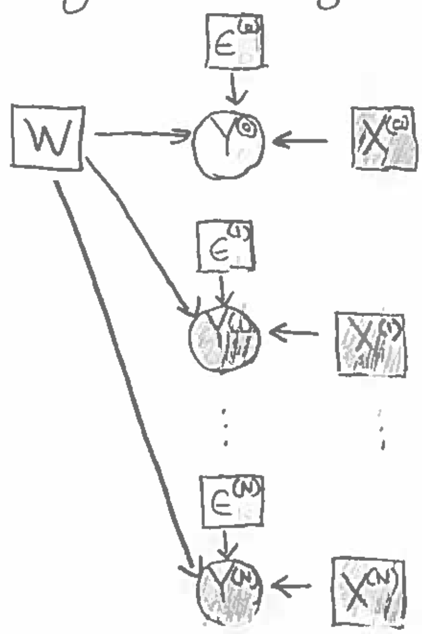
Linear Regression: MAP

① Recall "ordinary linear regression":



where: $P_\epsilon(\epsilon^{(n)}) \sim \text{Normal}(0, \sigma^2) \quad \forall n \in \{0, \ldots, N\}$

$y^{(n)} \leftarrow w^T x^{(n)} + \epsilon^{(n)}$

---

② Also recall that one way to estimate the value of the unobserved response variable $Y^{(0)}$ is through maximum a posteriori (MAP) estimation:

(a) compute $\hat{w} = \underset{w}{\text{argmax}}\ P(w) \prod_{n=1}^{N} P(y^{(n)} | w, x^{(n)})$

(b) compute $\hat{y}^{(0)} = \underset{y^{(0)}}{\text{argmax}}\ P(y^{(0)} | \hat{w}, x^{(0)})$

In the MLE approach, we assume that all weight vectors are equally likely (without further evidence), so we treat $P(w)$ as a constant and drop it from the equation.

LINEAR REGRESSION: MAP

(3) But maybe we do have an opinion about which weight vectors are more likely <u>prior to</u> observing any evidence (this is called a <u>prior probability</u> or an <u>apriori belief</u>).

First off, why would we have such an opinion?

(4) Consider if we actually wanted to predict someone's cholesterol accurately on the basis of lifestyle factors. We don't know what might be relevant, so we throw a lot of evidence variables into the mix:

X (evidence vars)					Y (response var)
X_1	X_2	X_3	X_4	X_{10000}	
(offset)	(age)	(weight)	(smoking freq)	(gumchewing freq)	(cholesterol)

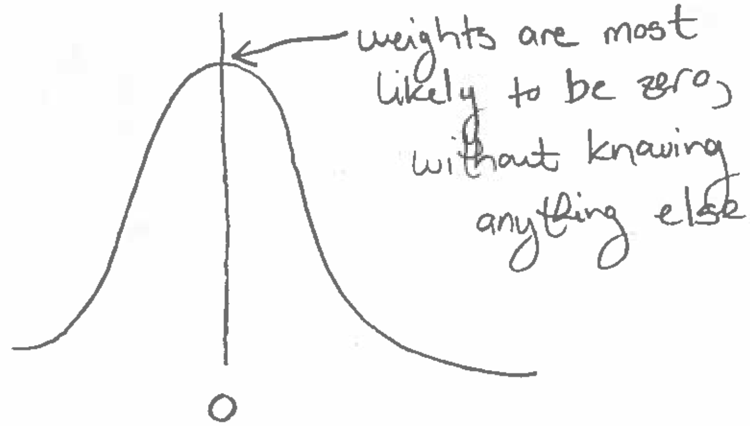
(5) Most of these evidence vars probably don't have any impact on cholesterol, so we expect that a good weight vector $w = \begin{bmatrix} w_1 \\ \vdots \\ w_{10000} \end{bmatrix}$ will contain mostly zeroes, since then $Y = w^T x$ will only be a function of a small subset of the evidence vars.

LINEAR REGRESSION: MAP

⑥ So our apriori belief is that weights are most likely to be zero.

How can we express this as a distribution?

One way is to say that for each weight w_d, $P(w_d) \sim \text{Normal}(0, \sigma^2)$ for some variance σ^2:



⑦ So let's see if we can simplify ②(a), our point estimate:

$$\hat{w} = \underset{w}{\operatorname{argmax}}\ P(w) \prod_{n=1}^{N} P(y^{(n)} | w, x^{(n)})$$

$$= \underset{w}{\operatorname{argmax}}\ \log P(w) \prod_{n=1}^{N} P(y^{(n)} | w, x^{(n)})$$

$$= \underset{w}{\operatorname{argmax}}\ \ell(w)$$

Linear Regression: MAP

⑧ Continuing:

$$\ell(w) = \log P(w) + \sum_{n=1}^{N} \log P(y^{(n)} | w, x^{(n)})$$

$$= \log P(w) + \ell_{MLE}(w)$$

where $\ell_{MLE}(w)$ is the likelihood function for the MLE (see Linear Regression: MLE, ④)

---

⑨ As with ordinary linear regression, we'll assume the stochastic terms $\epsilon^{(n)}$ are normally distributed, i.e.

$$P_\epsilon \sim \text{Normal}(0, \sigma^2)$$

We'll also use the prior distribution over weights that we argued for in ⑥:

$$P(w) \sim \text{Normal}(0, \tau^2)$$

not necessarily the same variance

These choices give us a type of regression called <u>ridge regression</u>.

LINEAR REGRESSION: MAP

(10) Continuing to simplify with these choices:

$$\hat{w} = \underset{w}{\operatorname{argmax}}\ \ell(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \log P(w) + \ell_{MLE}(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \log\left(\prod_{d=1}^{D}\left(\frac{1}{2\pi\tau^2}\right)^{\frac{1}{2}}\exp\left(\frac{-1}{2\tau^2}w_d^2\right)\right) + \ell_{MLE}(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \sum_{d=1}^{D}\log\left[\left(\frac{1}{2\pi\tau^2}\right)^{\frac{1}{2}}\exp\left(\frac{-1}{2\tau^2}w_d^2\right)\right] + \ell_{MLE}(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \left(\sum_{d=1}^{D}\frac{-1}{2}\log 2\pi\tau^2\right) + \left(\sum_{d=1}^{D}\frac{-1}{2\tau^2}w_d^2\right) + \ell_{MLE}(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \frac{-D}{2}\log 2\pi\tau^2 - \frac{1}{2\tau^2}\sum_{d=1}^{D}w_d^2 + \ell_{MLE}(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \frac{-\sum_{d=1}^{D}w_d^2}{2\tau^2} + \ell_{MLE}(w)$$

$$= \underset{w}{\operatorname{argmax}}\ \frac{-w^T w}{2\tau^2} + \ell_{MLE}(w)$$

LINEAR REGRESSION: MAP

(11) From LINEAR REGRESSION: MLE, (5), we know that:

$$\ell_{MLE}(w) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y^{(n)} - w^T x^{(n)}\right)^2$$

$$= -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)$$

Plugging this in to what we have so far:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \frac{-w^T w}{2\tau^2} - \cancel{\frac{N}{2}\log 2\pi\sigma^2} - \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)$$

$$= \underset{w}{\operatorname{argmax}} -\frac{1}{2\tau^2} w^T w - \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)$$

$$= \underset{w}{\operatorname{argmax}} -\frac{\sigma^2}{\tau^2} w^T w - (y - Xw)^T(y - Xw)$$

$= \ldots$ [See LINEAR REGRESSION: MLE (8)]

$$= \underset{w}{\operatorname{argmax}} -\frac{\sigma^2}{\tau^2} w^T w + 2w^T X^T y - w^T X^T X w$$

$$= \underset{w}{\operatorname{argmax}}\ 2w^T X^T y - w^T X^T X w - \frac{\sigma^2}{\tau^2} w^T w$$

$$= \underset{w}{\operatorname{argmin}}\ w^T X^T X w - 2w^T X^T y + \frac{\sigma^2}{\tau^2} w^T w$$

LINEAR REGRESSION: MAP

(12) So the loss function for ridge regression is:

$$L_{ridge}(w) = w^T X^T X w - 2w^T X^T y + \frac{\sigma^2}{\tau^2} w^T w$$

Notice that this can be expressed in terms of the loss function for ordinary linear regression $L_{lin}(w) = w^T X^T X w - 2w^T X^T y$:

$$L_{ridge}(w) = L_{lin}(w) + \frac{\sigma^2}{\tau^2} w^T w$$

(13) That means ridge regression's loss function is simply the usual linear regression loss function, plus some constant multiple of the squared "L_2-norm" of the weight vector:

$$w^T w = \sum_{d=1}^{D} w_d^2 = \|w\|_2^2$$

$$\hat{w} = \underset{w}{\operatorname{argmin}}\ L_{ridge}(w)$$

$$= \underset{w}{\operatorname{argmin}}\ \underbrace{L_{lin}(w)}_{\text{we want the likelihood of the data to be high}} + \underbrace{K \cdot w^T w}_{\text{but we also want the weights to be close to zero}}$$

(14) Thus ridge regression's loss function is combining two different objectives:

(a) we want the likelihood of the training data to be high $\qquad \underset{w}{\operatorname{argmin}} \; L_{lin}(w)$

(b) we want the learned weight vector to have a small L_2-norm (i.e. we want the length of the weight vector to be small) $\qquad \underset{w}{\operatorname{argmin}} \; \|w\|_2^2$

Objective (b) is often called a <u>regularization term</u> and so ridge regression is sometimes known as <u>linear regression with L_2-regularization</u>.

(15) So going back to (2), ridge regression estimates the value of the unobserved response variable $Y^{(0)}$ as follows:

(a) compute point estimate
$$\hat{w} = \underset{w}{\operatorname{argmax}} \; P(w) \prod_{n=1}^{N} P(y^{(n)} \mid w, x^{(n)})$$
$$= \underset{w}{\operatorname{argmin}} \; L_{ridge}(w)$$

(b) compute
$$\hat{y}^{(0)} = \underset{y^{(0)}}{\operatorname{argmax}} \; P(y^{(0)} \mid \hat{w}, x^{(0)})$$
$$= \hat{w}^T x^{(0)}$$
(see LINEAR REGRESSION: MLE, (3))

LINEAR REGRESSION: MAP

⑥ Exercise: Adapt LINEAR REGRESSION: MLE ⑩–⑪ to compute a closed-form expression for $\underset{w}{\operatorname{argmin}}\, L_{ridge}(w)$.