


REGRESSION PROBLEMS

① In a regression problem, we have a set of evidence variables X_1, \dots, X_D and a response variable Y that we want to predict.

For instance, let's say we want to predict cholesterol level given age and weight:

X (evidence vars)			Y (response var)	
	x_1 (offset)	x_2 (age)	x_3 (weight)	(cholesterol)
$x^{(1)}$	[1	24	150]	182 = $y^{(1)}$
$x^{(2)}$	[1	50	164]	210 = $y^{(2)}$
		\vdots		\vdots
$x^{(N)}$	[1	22	205]	202 = $y^{(N)}$

We have N training examples, each consisting of a vector $x^{(n)}$ and a scalar $y^{(n)}$, to learn from. Note that the entire dataset can be captured as an evidence matrix X and response vector y .

X (evidence matrix)

$$\begin{bmatrix} 1 & 24 & 150 \\ 1 & 50 & 164 \\ \vdots & \vdots & \vdots \\ 1 & 22 & 205 \end{bmatrix}$$

y (response vector)

$$\begin{bmatrix} 182 \\ 210 \\ \vdots \\ 202 \end{bmatrix}$$

REGRESSION PROBLEMS

② We assume the response variable is generated using the following steps:

- we start with a vector of weights (one weight per evidence variable), e.g.

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} -50 \\ 2 \\ 1 \end{bmatrix}$$

- we compute the weighted linear combination of the evidence variables (this will result in a single number):

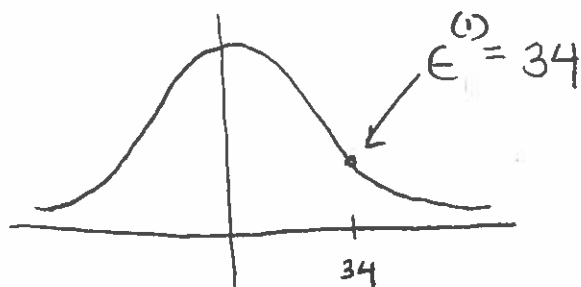
$$x^{(1)} w = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

$$= -50 \cdot 1 + 2 \cdot 24 + 1 \cdot 150 \\ = 148$$

the offset allows us to shift the total up or down by a constant factor



- we sample a random number $\epsilon^{(n)}$ from a distribution P_ϵ
e.g. $\epsilon^{(1)} \propto \text{Normal}(0, \sigma^2)$

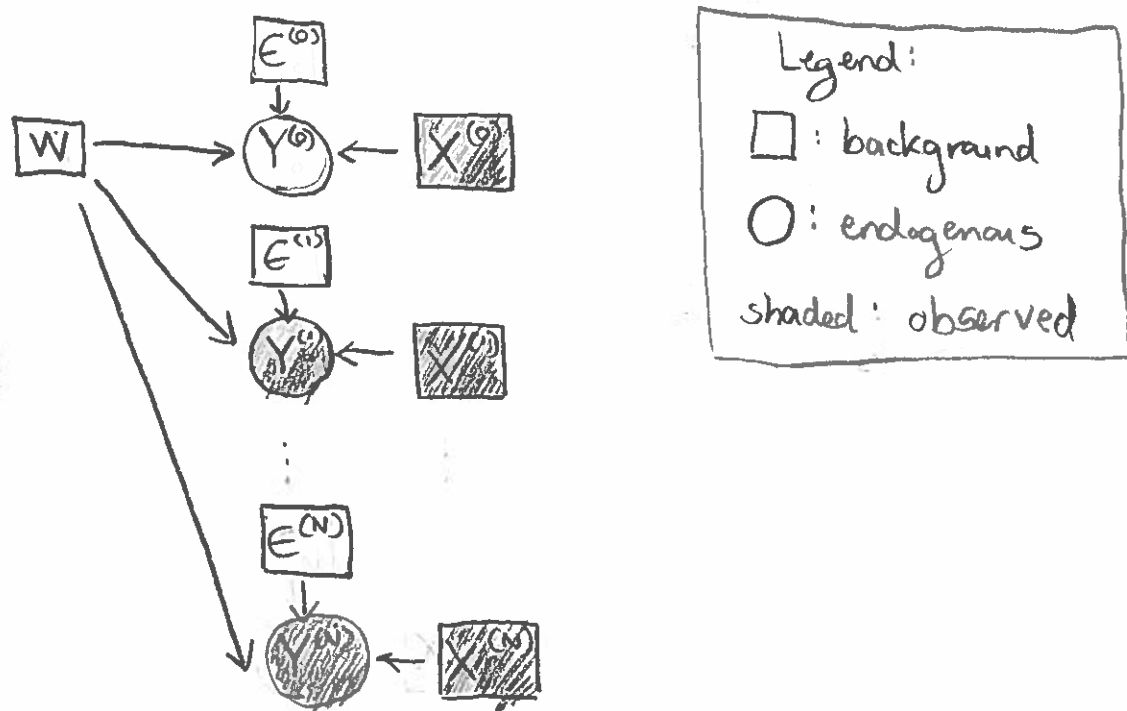


- we compute the response variable as a function p of $x^{(n)} w$ and $\epsilon^{(n)}$, e.g.

$$Y = p(x^{(1)} w, \epsilon^{(1)}) = x^{(1)} w + \epsilon^{(1)} = 148 + 34 = 182$$

REGRESSION PROBLEMS

- ③ Given this setup, we want to predict the value of an unobserved response variable $y^{(o)}$ given its observed evidence variables $x^{(o)}$ and our training data. In its simplest formulation, the causal diagram looks as follows:



We assume a probability distribution P over the background variables such that all background variables are marginally independent, i.e.

$$P(w, \epsilon^{(o)}, \dots, \epsilon^{(n)}, x^{(o)}, \dots, x^{(n)}) = P_w(w) P_\epsilon(\epsilon^{(o)}) \dots P_\epsilon(\epsilon^{(n)}) P_x(x^{(o)}) \dots P_x(x^{(n)})$$

Moreover we assume that all variables $\epsilon^{(n)}$ are drawn from the same distribution P_ϵ and all variables $x^{(n)}$ are drawn from the same distribution P_x .

REGRESSION PROBLEMS

④ What's the deal with the $\epsilon^{(n)}$'s?

We'll call these stochastic terms. The idea is to allow some softness around the deterministic point $x^{(n)}w$. In other words, just because $x^{(n)}w = 148$ for age = 24 and weight = 150, that doesn't mean we want the model to claim that EVERY person whose age is 24 and whose weight is 150 must have a cholesterol level of EXACTLY 148.

Instead, we want their cholesterol levels to be dispersed around 148.

⑤ Suppose we choose $P_\epsilon \sim \text{Normal}(0, \sigma^2)$ for some fixed variance σ^2 , and suppose we choose $p(z, \epsilon) = z + \epsilon$. This is what we did in ②, which gave us a cholesterol level of $148 + 34 = 182$ for a 24-year-old, 150lb subject.

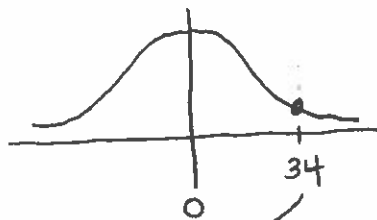
DETERMINISTIC TERM

$$w^T x = \begin{bmatrix} -50 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 24 \\ 150 \end{bmatrix}$$

$$= 148$$

STOCHASTIC TERM

$$\epsilon =$$



$$\oplus$$

$$182$$

REGRESSION PROBLEMS

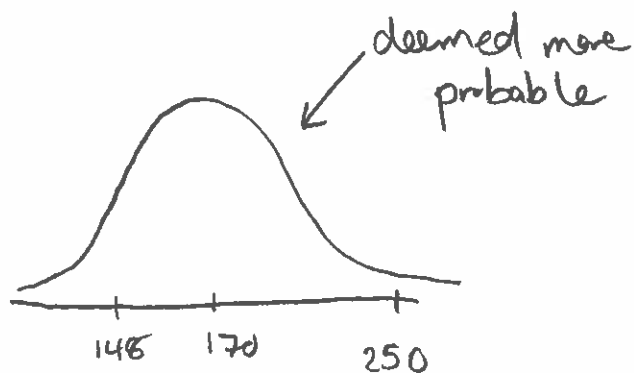
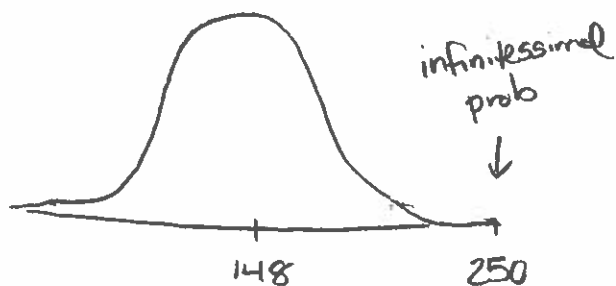
⑥ These choices:

$$\begin{cases} P_{\epsilon} \sim \text{Normal}(0, \sigma^2) \\ P(Z, \epsilon) = Z + \epsilon \end{cases}$$

give us the ordinary linear regression model

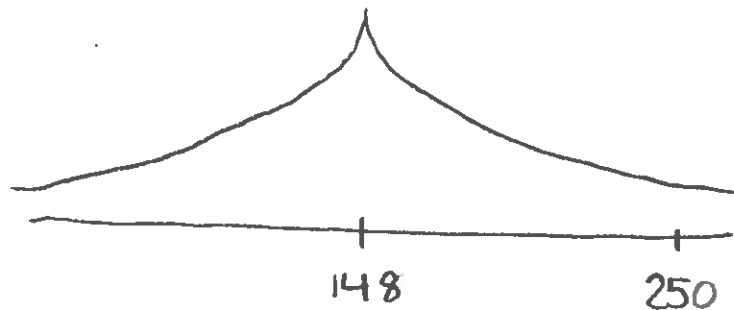
⑦ But it may not be the best choice. One possible downside of the normal distribution is that it has rapidly diminishing tails, so a normal distribution centered at 148 may give a nearly infinitesimal probability to 250.

In practice, this means that using the normal distribution makes regression very sensitive to outliers and noise, because even you have a thousand 24-year-olds with cholesterol 148, just one 24-year-old with cholesterol 250 can cause the learned mean to shift significantly upwards, because this will be deemed more probable than even one person with cholesterol 250:



REGRESSION PROBLEMS

- ⑧ If this is a problem for you, you can use a "heavy-tailed" distribution (i.e. they diminish in probability much more slowly). One example is the Laplace distribution.



Often, however, such distributions are not as computationally convenient.

-
- ⑨ This choice:

$$\begin{cases} P_{\epsilon} \sim \text{Laplace}(0, b) \\ \rho(z, \epsilon) = z + \epsilon \end{cases}$$

gives us the "robust" linear regression model

REGRESSION PROBLEMS

- ⑩ Another common choice for P_e and p comes into play when the response variable is Boolean-valued, e.g.:

X (evidence vars)			Y (response var)				
x_1	x_2	x_3					
(offset)	(age)	(weight)	(has high cholesterol)				
$x^{(1)} = [$	1	24	150	$]$	0	=	$y^{(1)}$
$x^{(2)} = [$	1	50	164	$]$	1	=	$y^{(2)}$
\vdots					\vdots		
$x^{(n)} = [$	1	22	205	$]$	1	=	$y^{(n)}$

- ⑪ We could use ordinary linear regression, but then we end up predicting values in the range $(-\infty, \infty)$, rather than restricting ourselves to the set $\{0, 1\}$.

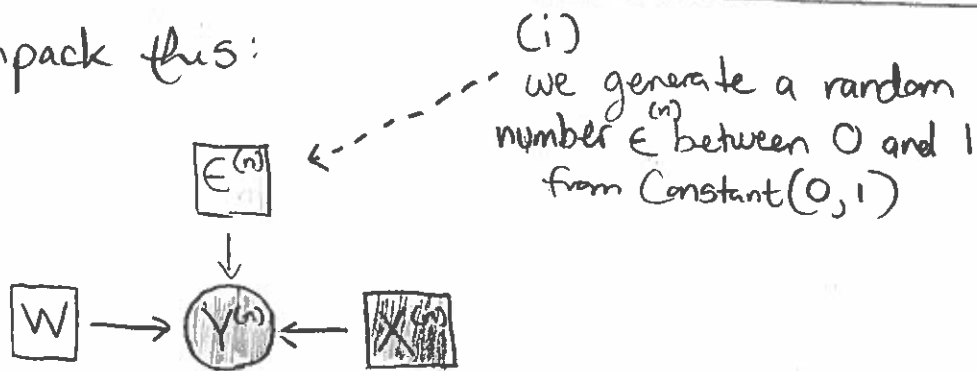
Another choice:

$$\begin{cases} P_e \sim \text{Uniform}(0, 1) \\ p(z, e) = \frac{1}{1 + e^{-z}}(z) = \begin{cases} 1 & \text{if } e < \frac{1}{1 + e^{-z}} \\ 0 & \text{o.w.} \end{cases} \end{cases}$$

gives us the famous logistic regression model

REGRESSION PROBLEMS

⑫ Let's unpack this:



(ii) we compute a threshold

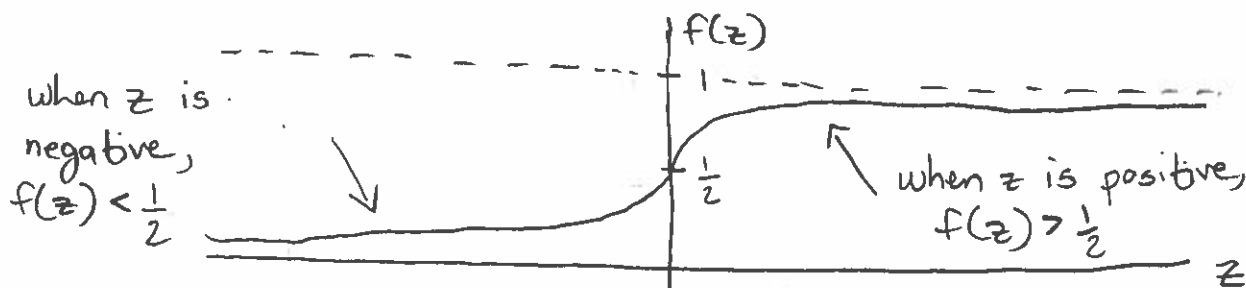
$$\tau = \frac{1}{1 + e^{-x^{(n)}w}}$$

(note: $0 < \tau < 1$)

(iii) $y^{(n)} = \begin{cases} 1 & \text{if } \epsilon < \tau \\ 0 & \text{otherwise} \end{cases}$

In other words, the probability that $y^{(n)} = 1$ is equal to $\frac{1}{1 + e^{-x^{(n)}w}}$

⑬ This function, $f(z) = (1 + e^{-z})^{-1}$, looks like this:



is called the logistic function (or the logit, or the sigmoid).

REGRESSION PROBLEMS

- ⑭ No matter what regression model we're using, we typically want to predict the value of the unobserved response variable $y^{(0)}$ given its observed evidence variables $x^{(0)}$ (recall $x^{(0)}$ is a vector) and our other observations (i.e. $x^{(n)}, y^{(n)}$ for $n \geq 1$).

In other words, we want
$$\operatorname{argmax}_{y^{(0)}} P(y^{(0)} | x^{(0)}, x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

- ⑮ The exact computation would be:

$$\begin{aligned} & \operatorname{argmax}_{y^{(0)}} P(y^{(0)} | x^{(0)}, x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) \\ &= \operatorname{argmax}_{y^{(0)}} \int P(y^{(0)}, w | x^{(0)}, x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) dw \\ & \quad \text{[Law of Total Probability]} \\ &= \operatorname{argmax}_{y^{(0)}} \int P(y^{(0)} | w, x^{(0)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) P(w | x^{(0)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) dw \\ & \quad \text{[Chain Rule]} \\ &= \operatorname{argmax}_{y^{(0)}} \int P(y^{(0)} | w, x^{(0)}) P(w | x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) dw \\ & \quad \text{[d-separation]} \end{aligned}$$

REGRESSION PROBLEMS

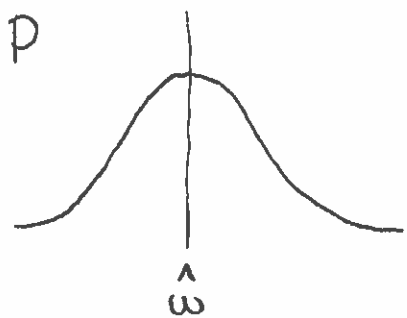
⑩ This integral is going to be messy, so let's make a simplifying assumption. Instead of using the actual distribution over weights:

$$P(w | x^{(1)}, \dots, x^{(n)}, y^{(1)}, \dots, y^{(n)})$$

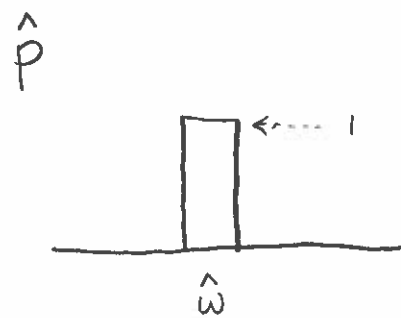
let's use a much simpler distribution:

$$\hat{P}(w | x^{(1)}, \dots, x^{(n)}, y^{(1)}, \dots, y^{(n)}) = \begin{cases} 1 & \text{if } w = \underset{w}{\operatorname{argmax}} P(w | x^{(1)}, \dots, y^{(n)}) \\ 0 & \text{o.w.} \end{cases}$$

This is called the point estimate approach, because it concentrates all probability mass onto the most likely value:



becomes
~>



REGRESSION PROBLEMS

①⑦ Let's make this approximation:

$$\operatorname{argmax}_{y^{(0)}} \int P(y^{(0)} | w, x^{(0)}) P(w | x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) dw$$

$$\approx \operatorname{argmax}_{y^{(0)}} \int P(y^{(0)} | w, x^{(0)}) \hat{P}(w | x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)}) dw$$

$$= \operatorname{argmax}_{y^{(0)}} P(y^{(0)} | \hat{w}, x^{(0)})$$

$$\text{where } \hat{w} = \operatorname{argmax}_w P(w | x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

①⑧ So the point estimate approach to predicting an unknown response $y^{(0)}$ has two steps:

(a) compute the most probable weight vector \hat{w} given the observations:

$$\hat{w} = \operatorname{argmax}_w P(w | x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

(b) compute the most probable response $\hat{y}^{(0)}$ given \hat{w} and evidence $x^{(0)}$:

$$\hat{y}^{(0)} = \operatorname{argmax}_{y^{(0)}} P(y^{(0)} | \hat{w}, x^{(0)})$$

REGRESSION PROBLEMS

① Part (a) can be simplified:

$$\hat{w} = \operatorname{argmax}_w P(w | x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

$$= \operatorname{argmax}_w \frac{P(x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)} | w) P(w)}{P(x^{(1)}, \dots, y^{(N)})} \quad [\text{Bayes Rule}]$$

$$= \operatorname{argmax}_w P(x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)} | w) P(w) \quad [\text{remove constant factor from argmax}]$$

$$= \operatorname{argmax}_w P(x^{(1)} | w) P(x^{(2)} | x^{(1)}, w) \dots P(y^{(N)} | x^{(1)}, \dots, y^{(N-1)}, w) P(w) \quad [\text{Chain Rule}]$$

$$= \operatorname{argmax}_w P(x^{(1)}) P(x^{(2)}) \dots P(x^{(N)}) P(y^{(1)} | w, x^{(1)}) \dots P(y^{(N)} | w, x^{(N)}) \cdot P(w) \quad [d\text{-sep} \rightarrow \text{see } \textcircled{3}]$$

$$= \operatorname{argmax}_w P(y^{(1)} | w, x^{(1)}) \dots P(y^{(N)} | w, x^{(N)}) P(w) \quad [\text{remove constant factors}]$$

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

REGRESSION PROBLEMS

② In short:

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

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

This is called the MAP (maximum a posteriori) estimate.

② A special case of the MAP estimate assumes that $P(w)$ is the same for every possible w . Since it then becomes a constant factor, we can drop it from the argmax :

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

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

This is called the MLE (maximum likelihood estimate).

how can this be, if there's an infinite space of w -vectors? don't worry about it

