REGRESSION PROBLEMS

① In a regression problem, we have a set of evidence variables $X_1, \ldots, X_D$ 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:

often called the "bias"

	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	=	$Y^{(2)}$
⋮		⋮				⋮
$X^{(3)} =$	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) y (response vector)

$$X = \begin{bmatrix} 1 & 24 & 150 \\ 1 & 50 & 164 \\ \vdots & \vdots & \vdots \\ 1 & 22 & 205 \end{bmatrix} \qquad y = \begin{bmatrix} 182 \\ 210 \\ \vdots \\ 202 \end{bmatrix}$$

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):

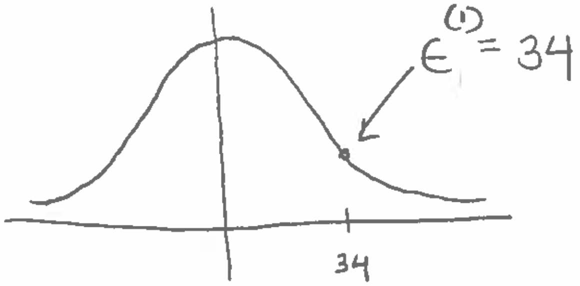
$$w^T X^{(1)} = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix}$$

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

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

$$= 148$$

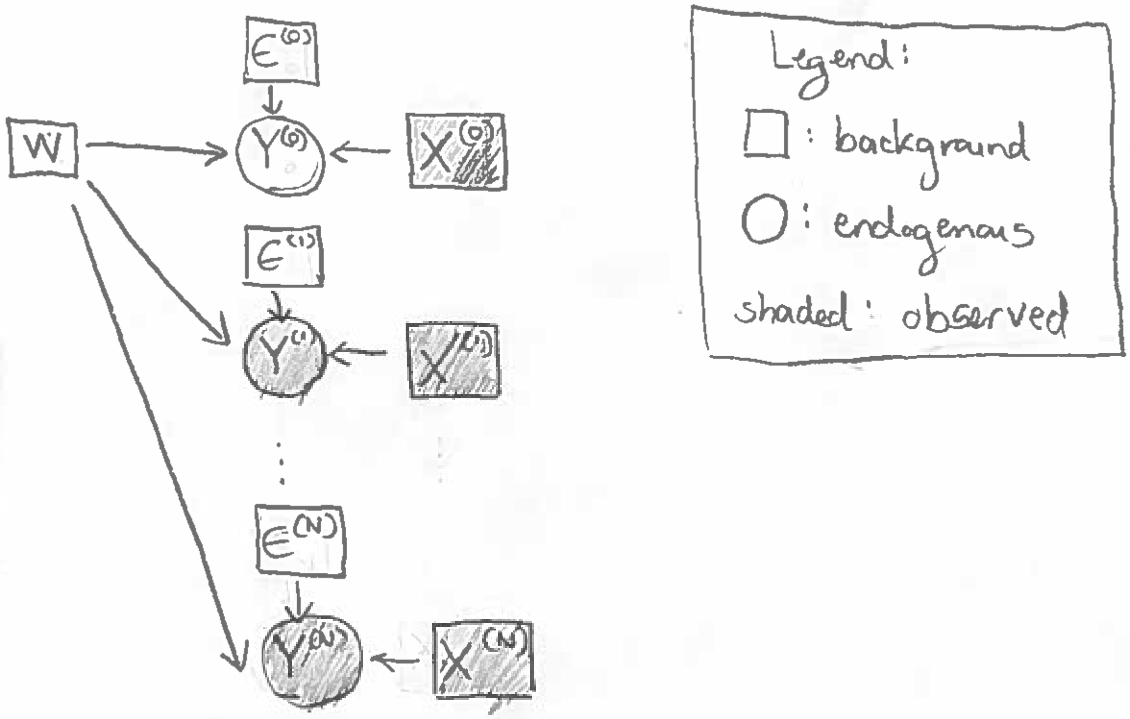
- we sample a random number ϵ from a distribution φ e.g. $\epsilon^{(1)} \propto \text{Normal}(0, \sigma^2)$



- we compute the response variable as a function ρ of of $w^T X$ and ϵ, e.g.

$$Y = \rho(w^T X^{(1)}, \epsilon^{(1)}) = w_T X^{(1)} + \epsilon^{(1)} = 148 + 34 = 182$$

③ Given this setup, we want to predict the value of an unobserved response variable $Y^{(0)}$ given its observed evidence variables $X^{(0)}$ 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.

$$P(w,\epsilon^{(0)},\ldots,\epsilon^{(N)},X^{(0)},\ldots,X^{(N)}) = P_w(w)P_\epsilon(\epsilon^{(0)})\ldots P_\epsilon(\epsilon^{(N)})\cdot P_x(X^{(0)})\ldots P(X$$

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.

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 $w^T x^{(n)}$. In other words, just because $w^T x^{(n)} = 148$ 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 $\rho(z, \epsilon) = z + \epsilon$. This is what we did in ②, which gave us a cholesterol level of $148 + 34 = 182$ for a 24-year-old, 150lb subject.

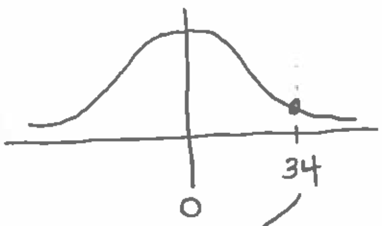
DETERMINISTIC TERM

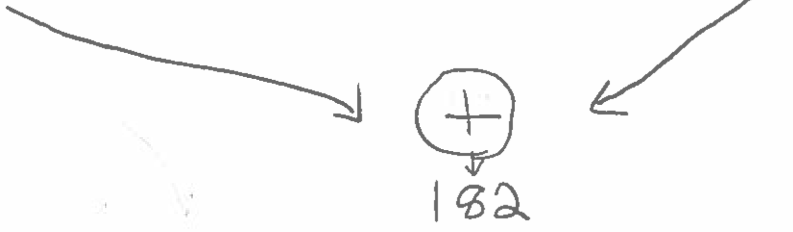
$$w^T x = \begin{bmatrix} -50 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 24 \\ 150 \end{bmatrix}$$

$$= 148$$

STOCHASTIC TERM

$\epsilon =$





REGRESSION PROBLEMS

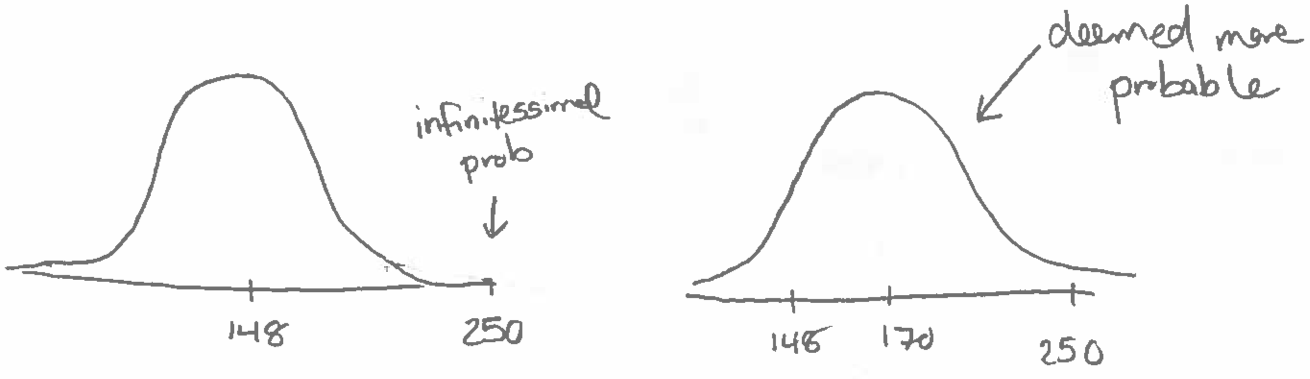
⑥ These choices:

$$\left[\begin{array}{l} P_\epsilon \sim \text{Normal}(0, \sigma^2) \\ \rho(z, \epsilon) = z + \epsilon \end{array}\right.$$

give us the ordinary linear regression model

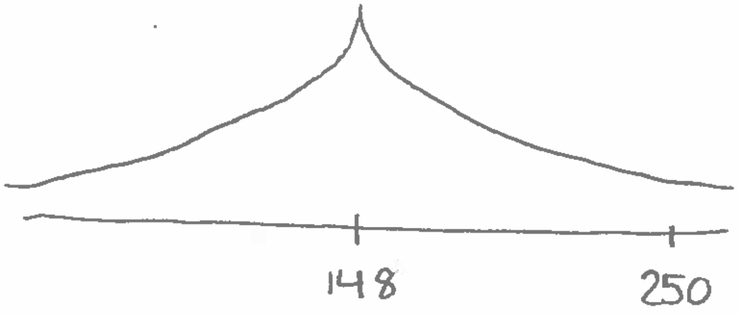
⑦ But it may not be the best choice. One possible downside of the normal distribution is that it has rapidly diminishing tails, so a normal distribution centered at 148 may give a nearly infinitessimal probability to 250.

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:

$$\left[\begin{array}{l} P_\epsilon \sim \text{Laplace}(0, b) \\ \rho(z, \epsilon) = z + \epsilon \end{array}\right.$$

gives us the "robust" linear regression model

(10) Another common choice for P_ϵ and ρ comes into play when the response variable is Boolean-valued, e.g.:

$$\begin{array}{rccc|ccc} & & X \text{ (evidence vars)} & & & Y \text{ (response var)} & \\ & X_1 & X_2 & X_3 & & & \\ & (\text{offset}) & (\text{age}) & (\text{weight}) & (\text{has high cholesterol}) & & \\ X^{(1)} = [& 1 & 24 & 150\ ] & 0 & = & Y^{(1)} \\ X^{(2)} = [& 1 & 50 & 164\ ] & 1 & = & Y^{(2)} \\ \vdots & & \vdots & & \vdots & & \vdots \\ X^{(N)} = [& 1 & 22 & 205\ ] & 1 & = & Y^{(N)} \end{array}$$

(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 $\{0, 1\}$.

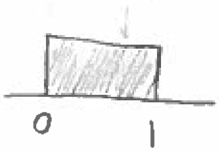
Another choice:

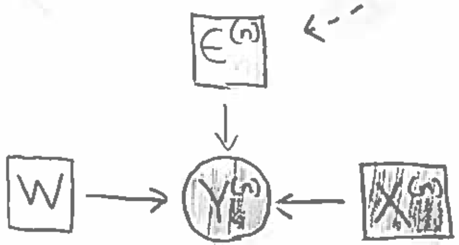
$$\left[\begin{array}{l} P_\epsilon \sim \text{Constant}(0,1) \\ \rho(z, \epsilon) = \mathbb{1}_{\epsilon < (1+e^{-z})^{-1}}(z) = \begin{cases} 1 & \text{if } \epsilon < \frac{1}{1+e^{-z}} \\ 0 & \text{o.w.} \end{cases} \end{array} \right.$$

gives us the famous <u>logistic regression</u> model

(12) Let's unpack this:

(i) we generate a random number $\epsilon^{(n)}$ between 0 and 1 from Constant(0,1)





(ii) we compute a threshold

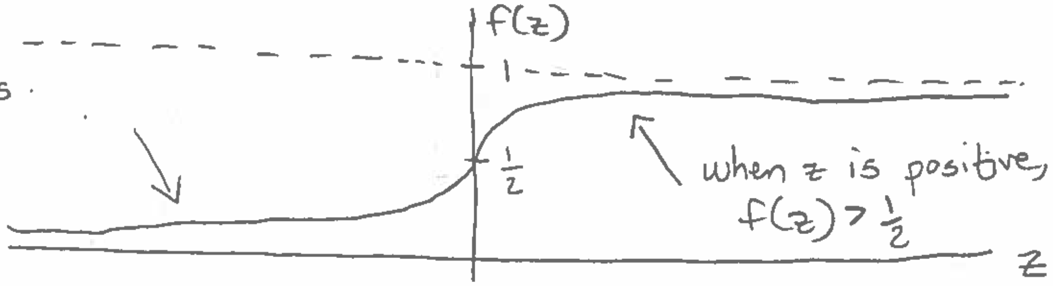
$$\tau = \frac{1}{1+e^{-w^T x^{(n)}}}$$

(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 $\dfrac{1}{1+e^{-w^T x^{(n)}}}$

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



when z is negative, $f(z) < \frac{1}{2}$

when z is positive, $f(z) > \frac{1}{2}$

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 $\underset{y^{(0)}}{\operatorname{argmax}} P(y^{(0)} | x^{(0)}, x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$

(15) The exact computation would be:

$$\underset{y^{(0)}}{\operatorname{argmax}} \; P(y^{(0)} | x^{(0)}, x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

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

[Law of Total Probability]

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

[Chain Rule]

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

[d-separation]

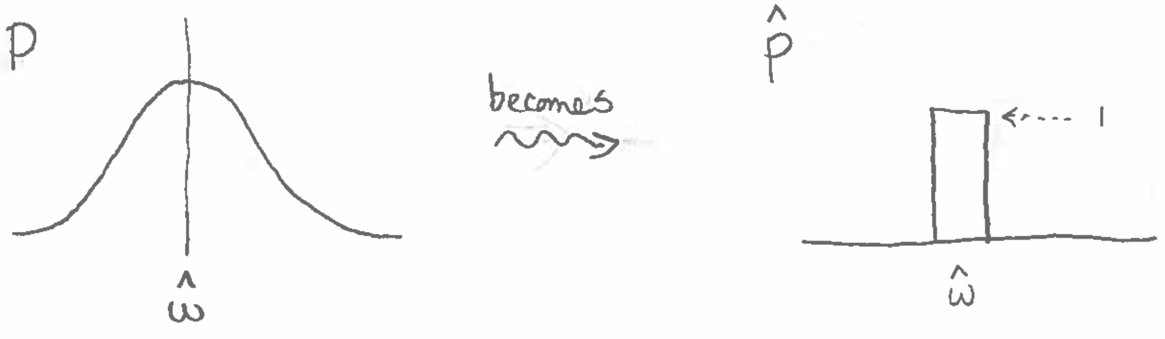
(16) This integral is going to be messy, so let's make a simplifying assumption. Instead of using the actual distribution over weights:

$$P(w \mid x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

let's use a much simpler distribution:

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

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



Regression Problems

(17) Let's make this approximation:

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

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

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

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

(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:

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

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

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

REGRESSION PROBLEMS

19 Part (a) can be simplified:

$$\hat{\omega} = \underset{\omega}{\operatorname{argmax}}\ P(\omega \mid x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)})$$

$$= \underset{\omega}{\operatorname{argmax}}\ \frac{P(x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)} \mid \omega) P(\omega)}{P(x^{(1)}, \dots, y^{(N)})}$$

[Bayes Rule]

$$= \underset{\omega}{\operatorname{argmax}}\ P(x^{(1)}, \dots, x^{(N)}, y^{(1)}, \dots, y^{(N)} \mid \omega) P(\omega)$$

[remove constant factor from argmax]

$$= \underset{\omega}{\operatorname{argmax}}\ P(x^{(1)} \mid \omega) P(x^{(2)} \mid x^{(1)}, \omega) \cdot \dots \cdot P(y^{(N)} \mid x^{(1)}, \dots, y^{(N-1)}, \omega) P(\omega)$$

[Chain Rule]

$$= \underset{\omega}{\operatorname{argmax}}\ P(x^{(1)}) P(x^{(2)}) \cdots P(x^{(N)}) P(y^{(1)} \mid \omega, x^{(1)}) \cdots P(y^{(N)} \mid \omega, x^{(N)}) \cdot P(\omega)$$

[d-sep → see 3]

$$= \underset{\omega}{\operatorname{argmax}}\ P(y^{(1)} \mid \omega, x^{(1)}) \cdot \dots \cdot P(y^{(N)} \mid \omega, x^{(N)}) P(\omega)$$

[remove constant factors]

$$= \underset{\omega}{\operatorname{argmax}}\ P(\omega) \prod_{n=1}^{N} P(y^{(n)} \mid \omega, x^{(n)})$$

REGRESSION PROBLEMS

20 In short:

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

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

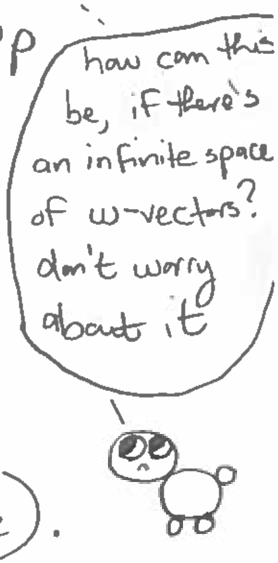
This is called the MAP (maximum a posteriori) estimate.

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

(a) compute $\hat{\omega} = \underset{\omega}{\operatorname{argmax}} \prod_{n=1}^{N} P(y^{(0)} | \omega, x^{(0)})$

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

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