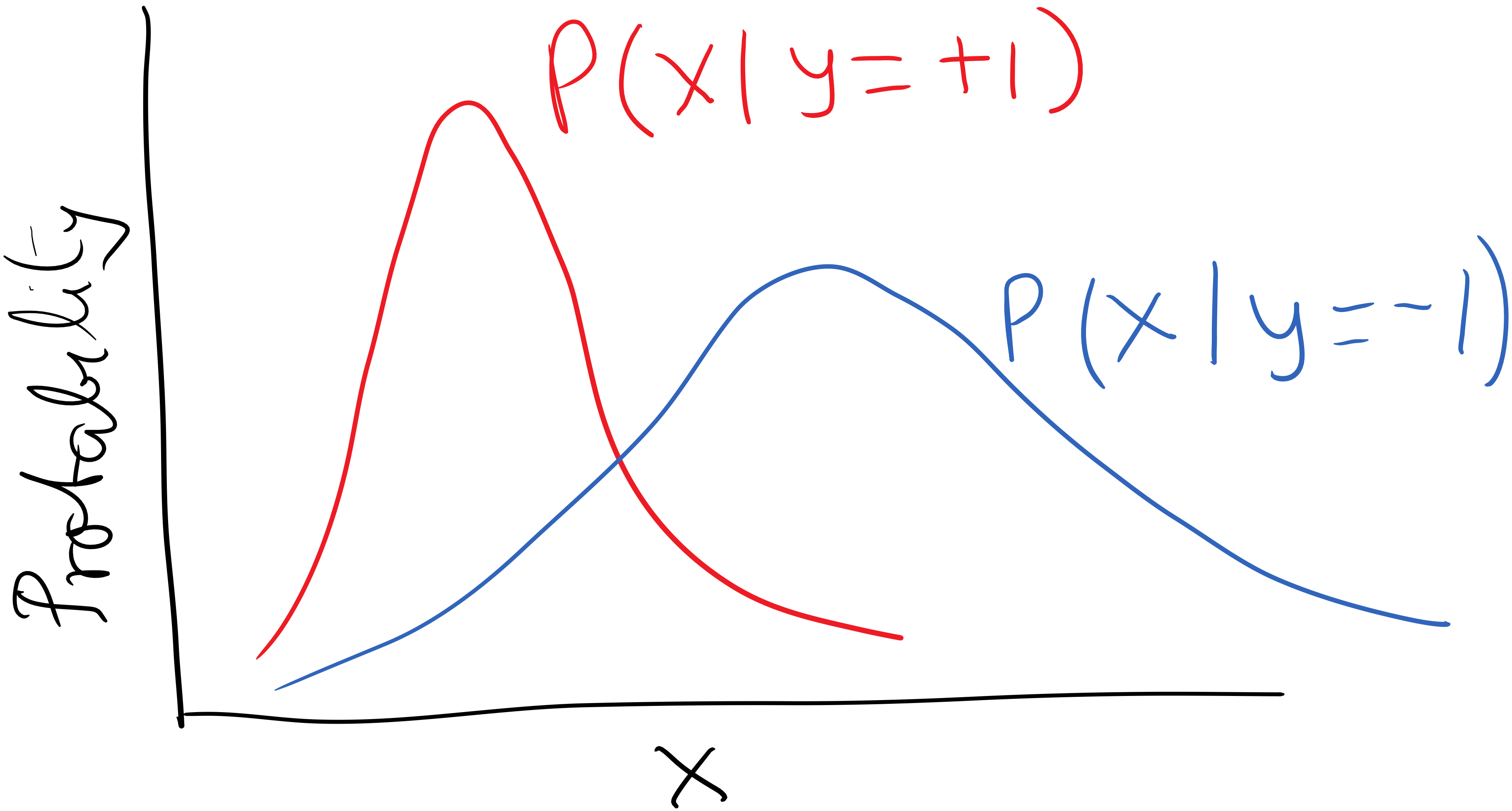
Classification

- Output is not a real number, but a label e.g. for an email, spam or not-spam
- Sometimes the label is binary, but it could be from a finite set e.g. {dog, cat, horse, rabbit}

Given x, what should we guess for y



x might be blood cholesterol and y=1 if healthy, -1 if heart disease

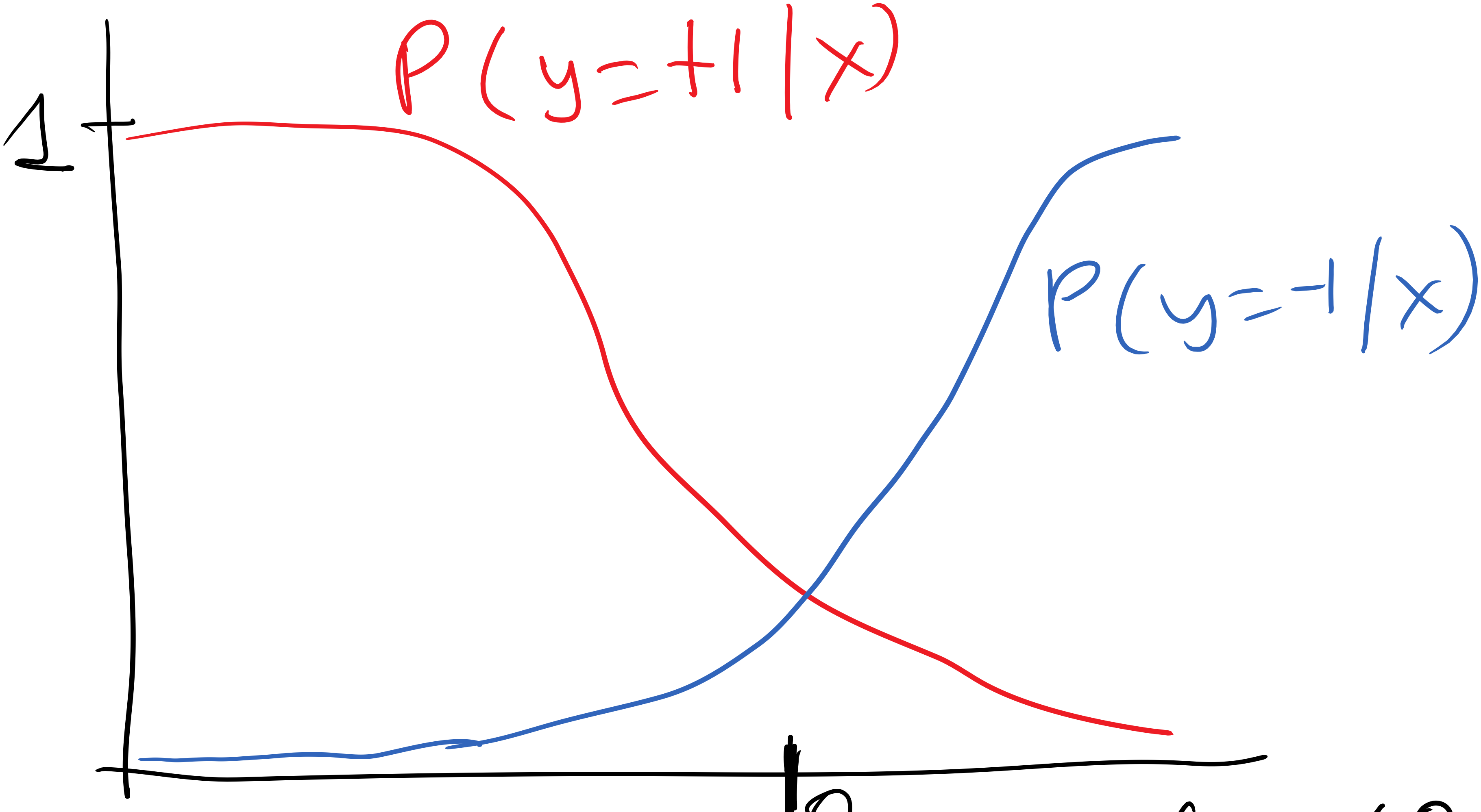
Use Bayes Rule

$$P(y=+1 \mid x) = \frac{P(x \mid y=+1)\,P(y=1)}{P(x)}$$

$$P(y=-1 \mid x) = \frac{P(x \mid y=-1)\,P(y=-1)}{P(x)}$$

Suppose $P(y=1) = 2/3$; $P(y=-1) = \frac{1}{3}$

Posterior probabilities



So should we guess $y = +1$ for $x < \theta$
-1 for $x > \theta$

Depends on the loss function!

- Yes, if the goal is to minimize the probability of misclassification

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error} \mid x)\, P(x)\, dx$$

To minimize $P(\text{error} \mid x)$ choose class with higher posterior probability

Three ways of building classifiers

- Generative
  - Model $P(\underset{\sim}{x} | C_K)$
  - Model $P(C_K)$

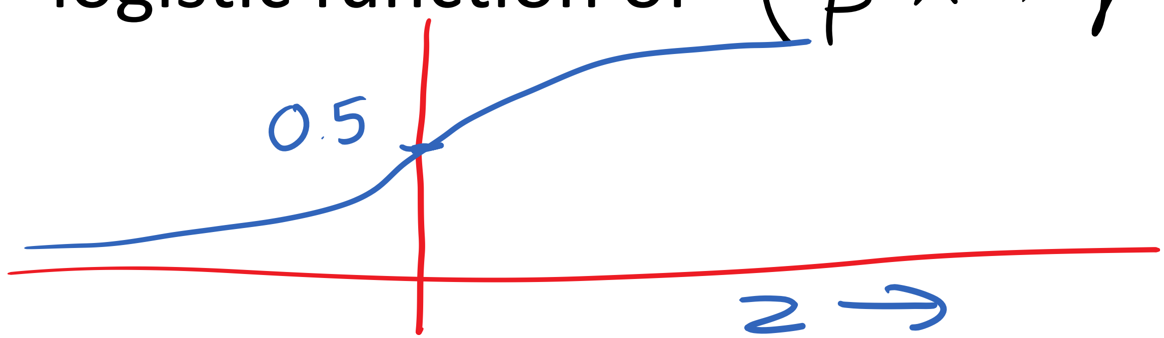
  obtain $P(C_K | \underset{\sim}{x})$ using Bayes Theorem
- Discriminative
  - Model $P(C_K | \underset{\sim}{x})$
- Find decision boundaries
  - Model $f : \underset{\sim}{x} \rightarrow K$

Logistic Regression

- Logistic Regression is a classification method (the output is binary, not a real number)
- We model the posterior probability by a logistic function whose argument is a linear function of the features + a bias term
- We can justify this choice in multiple ways
- Logistic regression can be extended to K classes, and is a building block for understanding neural networks.

Posterior probability for Gaussian class-conditional densities

- (important)We assume that the variances are the same for the different classes
- We do the calculation for 1D feature vectors initially; later we will see that the d-dimensional case works out quite similarly
- We will find that the posterior probability is a logistic function of $(\beta x + \gamma)$



$$\frac{1}{1+e^{-z}}$$

$$z = \beta x + \gamma$$

Proof that the posterior is a logistic(1)

$$P(x|C_1) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ \frac{-(x-\mu_1)^2}{2\sigma^2} \right\}$$

$$P(x|C_2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ \frac{-(x-\mu_2)^2}{2\sigma^2} \right\}$$

$$P(C_1|x) = \frac{P(x|C_1)\,P(C_1)}{P(x|C_1)\,P(C_1) + P(x|C_2)\,P(C_2)}$$

Substitute the expressions for

$P(x|C_1)$, $P(x|C_2)$, $P(C_1)=\pi_1$, $P(C_2)=1-\pi_1$

Proof that the posterior is a logistic(2)

$$P(C_1|x) = \frac{\pi_1 \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}}{\pi_1 \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\} + (1-\pi_1)\exp\left\{\frac{-(x-\mu_2)^2}{2\sigma^2}\right\}}$$

$$= \frac{1}{1 + \frac{1-\pi_1}{\pi_1}\exp\left\{\frac{-1}{2\sigma^2}\left[(x-\mu_2)^2 - (x-\mu_1)^2\right]\right\}}$$

$$= \frac{1}{1 + \frac{1-\pi_1}{\pi_1}\exp\left\{\frac{-1}{2\sigma^2}\left[2(\mu_1-\mu_2)x + (\mu_2^2-\mu_1^2)\right]\right\}}$$

Proof that the posterior is a logistic(3)

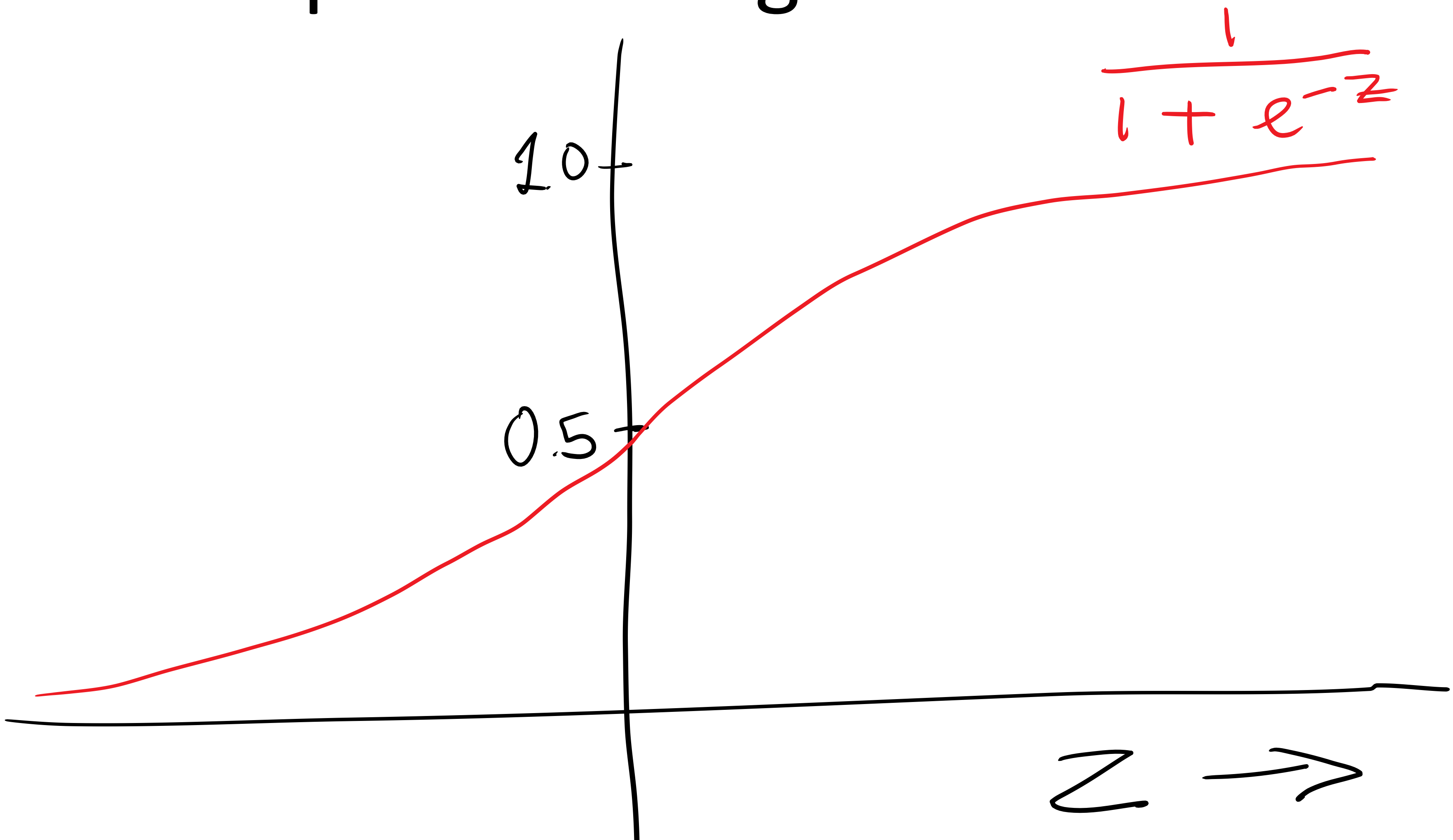
$$P(C_1|x) = \frac{1}{1 + \exp(-z)}$$

where $z = \beta x + \gamma$

$$\beta = \frac{\mu_1 - \mu_2}{\sigma^2}$$

$$\gamma = \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} + \ln\left(\frac{\pi_1}{1-\pi_1}\right)$$

Graph of the logistic function



z is a linear function of x + a bias term
(sometimes called "affine" function)

The posterior is logistic for multivariate Gaussians (1)

- Let us do the Multivariate case

$$\ln \frac{p(C_2|x)}{p(C_1|x)}$$

$$= -\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \ln\frac{\pi_2}{\pi_1}$$

$$= \mu_2^T \Sigma^{-1} x - \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 - \mu_1^T \Sigma^{-1} x + \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \ln\frac{\pi_2}{\pi_1}$$

$$= \beta^T x + \alpha$$

The posterior is logistic for multivariate Gaussians (2)

We know that ~~p~~ $p(C_1|x) = 1 - p(C_2|x)$

$$\ln \frac{p}{1-p} = \beta^T x + \alpha$$

This is sometimes called log-odds or logit

$$\frac{p}{1-p} = e^{\beta^T x + \alpha}$$

$$p = e^{\beta^T x + \alpha} - p\, e^{\beta^T x + \alpha}$$

$$p = \frac{e^{\beta^T x + \alpha}}{1 + e^{\beta^T x + \alpha}}$$

or $p(\underset{\sim}{x}) = \dfrac{1}{1 + e^{-(\beta^T x + \alpha)}}$

A heuristic argument for the logistic

We like linear models, but can we do so for probability?

(No, because $0 \leq p \leq 1$)

Can we do so for odds $\frac{p}{1-p}$

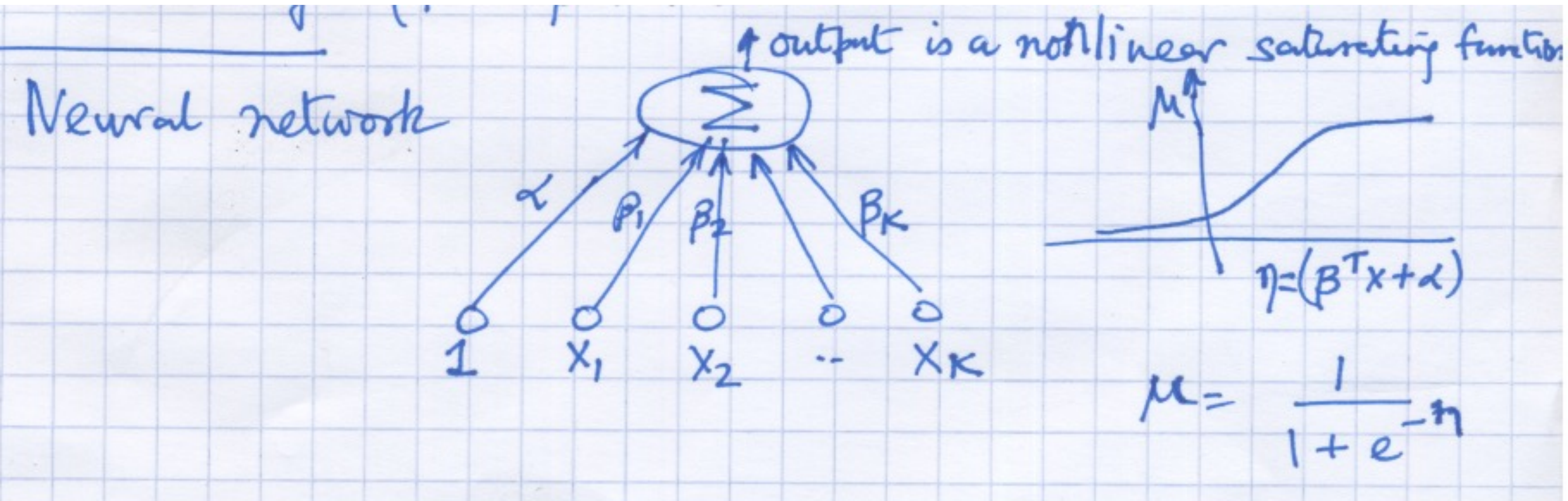
(better, because $(0, \infty)$ is range, but not symmetric)

Let's take logarithms! $\ln \frac{p}{1-p}$ (logit (p))

and now we have a linear model for the log-odds

$$\text{logit}(p) = \beta^T x + \alpha.$$

Neural networks can be modeled by logistics



Standard Trick: Add a 0^{th} component to the $\underset{\sim}{x}$ vector. which is fixed to be 1. This is connected with weight α

Modeling the probability distribution

We say that the class label Y is a Bernoulli random variable, with its probability parameter p being as above

$$P(Y=1 \mid x) = \frac{1}{1+\exp(-\beta^T x)}$$

For compactness, introduce notation $\mu(x) = \dfrac{1}{1+\exp(-\beta^T x)}$

or $\mu(x) = \dfrac{1}{1+\exp(-\eta(x))}$

$$P(Y=1 \mid x) = \mu(x)$$

As usual we use y to denote values taken by random variables

$$P(y \mid x) = \mu(x)^y \, (1-\mu(x))^{1-y}$$

Maximum Likelihood Estimation

Let's compute the likelihood

This is β in $\frac{1}{1+\exp(-\beta^T x)}$

$$P(y_1, \dots y_n \mid x_1, \dots x_n, \theta) = \prod_{i=1}^{n} \mu_i^{y_i} (1-\mu_i)^{(1-y_i)}$$

and the log-likelihood

$$\ell(\theta \mid D) = \sum_i y_i \ln \mu_i + (1-y_i) \ln (1-\mu_i)$$

We need to maximize this w.r.t. θ (β in this case)

Switching notation to make this explicit

$$\ell(\beta) = \sum_i y_i \ln \mu_i + (1-y_i) \ln (1-\mu_i)$$

Next, we compute the gradient of the log-likelihood

Compute gradient with respect to β

$$\nabla_\beta \ell = \sum_i \frac{y_i}{\mu_i} \frac{d\mu_i}{d\beta} - \frac{(1-y_i)}{(1-\mu_i)} \frac{d\mu_i}{d\beta}$$

Derivative of a logistic function

We will show $\frac{d\mu}{d\underset{\sim}{\beta}} = \mu(1-\mu)\,\underset{\sim}{x}$

<u>Proof</u>

$$\mu = \frac{1}{1+e^{-\underset{\sim}{\beta}^T \underset{\sim}{x}}}$$

$$e^{\underset{\sim}{\beta}^T \underset{\sim}{x}} = \frac{\mu}{1-\mu}$$

$$\underset{\sim}{\beta}^T \underset{\sim}{x} = \ln\left(\frac{\mu}{1-\mu}\right)$$

$$\frac{d}{d\beta}\, \underset{\sim}{\beta}^T \underset{\sim}{x} = \frac{d}{d\beta} \ln\left(\frac{\mu}{1-\mu}\right)$$

Derivative of a logistic (contd.)

$$\frac{d}{d\underset{\sim}{\beta}} \underset{\sim}{\beta}^T \underset{\sim}{x} = \frac{d}{d\underset{\sim}{\beta}} \ln\left(\frac{\mu}{1-\mu}\right)$$

$$\underset{\sim}{x} = \frac{1-\mu}{\mu} \cdot \frac{(1-\mu)+\mu}{(1-\mu)^2} \cdot \frac{d\mu}{d\underset{\sim}{\beta}}$$

$$\underset{\sim}{x} = \frac{1-\mu}{\mu} \cdot \frac{1}{(1-\mu)^2} \cdot \frac{d\mu}{d\underset{\sim}{\beta}}$$

$$\underset{\sim}{x} = \frac{1}{\mu(1-\mu)} \frac{d\mu}{d\underset{\sim}{\beta}}$$

$$\frac{d\mu}{d\underset{\sim}{\beta}} = \mu(1-\mu)\underset{\sim}{x}$$

Compute gradient with respect to β

$$\nabla_\beta l = \sum_i \frac{y_i}{\mu_i} \frac{d\mu_i}{d\beta} - \frac{(1-y_i)}{(1-\mu_i)} \frac{d\mu_i}{d\beta}$$

$$\nabla_\beta l = \sum_i \left\{ \frac{y_i}{\mu_i} - \frac{(1-y_i)}{(1-\mu_i)} \right\} \mu_i (1-\mu_i)\, x_i$$

$$= \sum_i \left\{ \frac{y_i - \cancel{y_i \mu_i} - \mu_i + \cancel{\mu_i y_i}}{\cancel{\mu_i (1-\mu_i)}} \right\} \cancel{\mu_i (1-\mu_i)}\, x_i$$

$$= \sum_i (y_i - \mu_i)\, x_i$$

In vector notation

$$\nabla_\beta l = X^T (\underset{\sim}{y} - \underset{\sim}{\mu})$$

$$X^T = \begin{vmatrix} x_1 & \cdots & x_n \end{vmatrix}$$

Stochastic Gradient Descent

If we want to increase the likelihood we take a step in the direction that will increase the likelihood

learning rate parameter

$$\beta^{(t+1)} = \beta^{(t)} + \rho \left(y_i - \mu_i^{(t)} \right) x_i$$