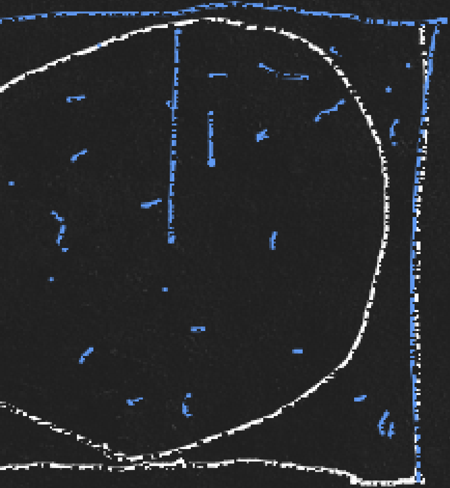
NOISE CONTRASTIVE ESTIMATION

OUTLINE

- Rejection Sampling
- Logistic Regression
- Unsupervised as Supervised Learning
- Noise Contrastive Estimation
- Word 2 vec and negative sampling

REJECTION SAMPLING

1



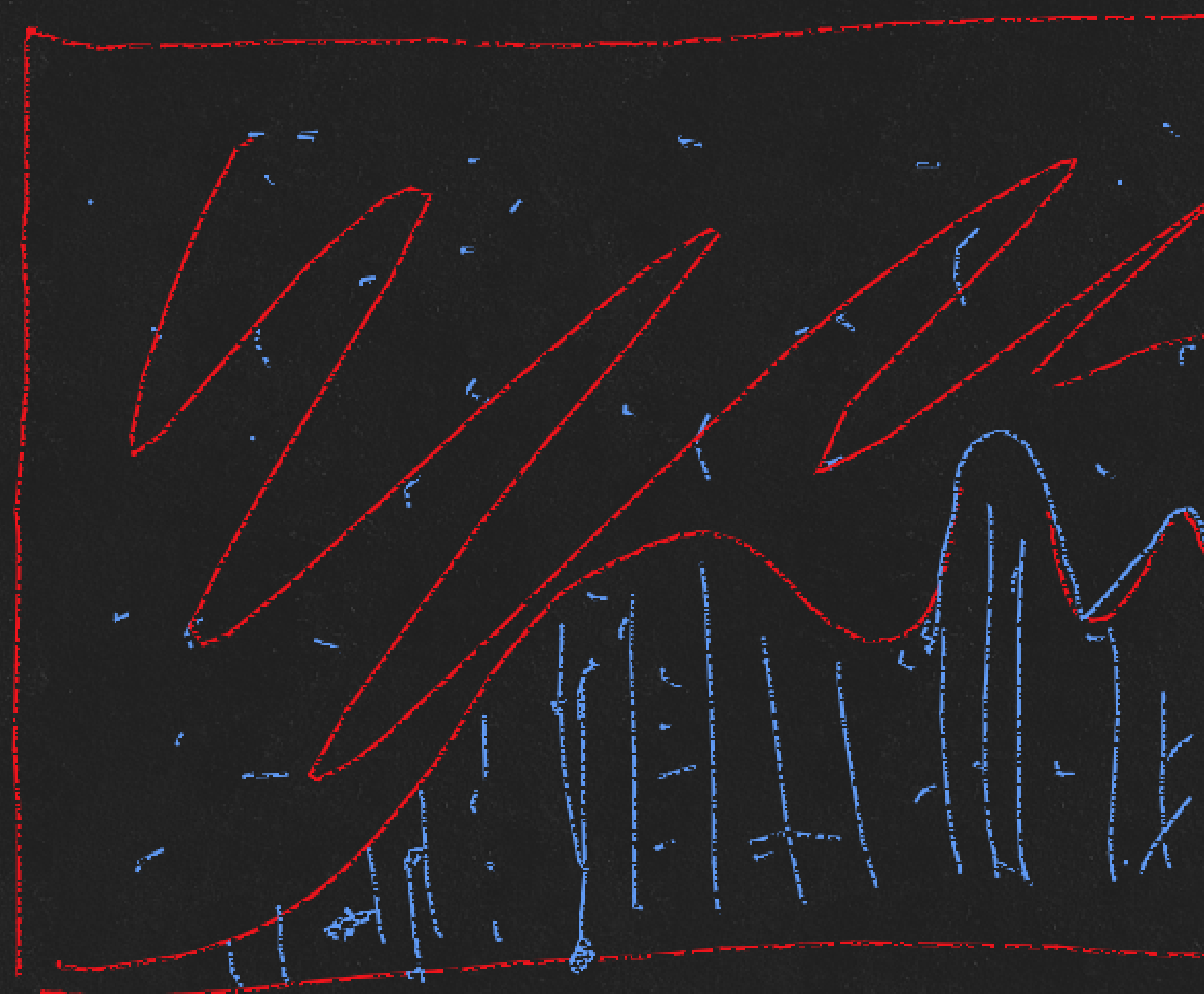
$$\text{area} = \pi r^2$$

$$\pi = \frac{\text{area}}{r^2}$$

1.

2.

3. $u > g(x)$ reject



LOGISTIC REGRESSION

PROBLEM DEFINITION

Logistic regression seeks to

- *Model* the probability of an event occuring depending on the values of the independent variables, which can be categorical or numerical
- *Estimate* the probability that an event occurs for a randomly selected observation versus the probability that the event does not occur
- *Predict* the effect of a series of variables on a binary response variable
- *Classify* observations by estimating the probability that an observation is in a particular category (e.g. approved or not approved for a loan)

OUR DATA IN 1D


1
0

BUT ALL I CAN DO IS FIT LINES?!

ODDS

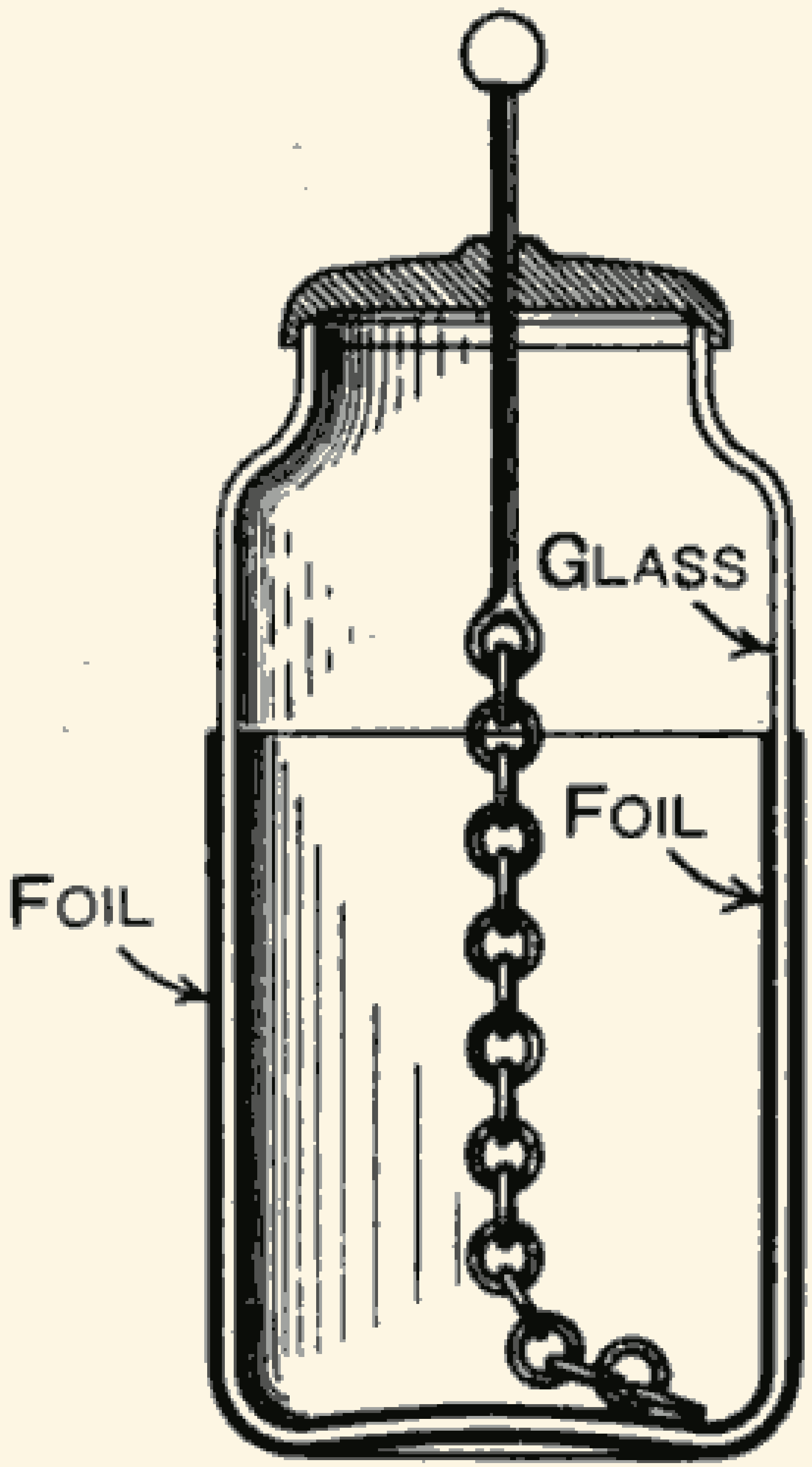
$$\frac{p_+}{1 - p_+}$$

Probability	Corresponding odds
0.5	50:50 or 1
0.9	90:10 or 9
0.999	999:1 or 999
0.01	1:99 or 0.0101
0.001	1:999 or 0.001001

LOG-ODDS

$$\log\left(\frac{p_+}{1 - p_+}\right)$$

Log-odds	Probability
0	0.5
2.19	0.9
6.9	0.999
-4.6	0.01
-6.9	0.001


GLASS
FOIL
FOIL

LINEAR FIT TO LOG-ODDS

$$
\begin{aligned}
\log\left(\frac{p_+}{1-p_+}\right) &= kx + b \\
&= w_1 x + w_0 \\
&= \mathbf{w}^T \mathbf{x}
\end{aligned}
$$

WHAT'S THE PROBABILITY?

$$\log\left(\frac{p_+}{1-p_+}\right) = \mathbf{w}^T\mathbf{x}$$

$$\frac{p_+}{1-p_+} = e^{\mathbf{w}^T\mathbf{x}}$$

$$p_+ = e^{\mathbf{w}^T\mathbf{x}}(1 - p_+)$$

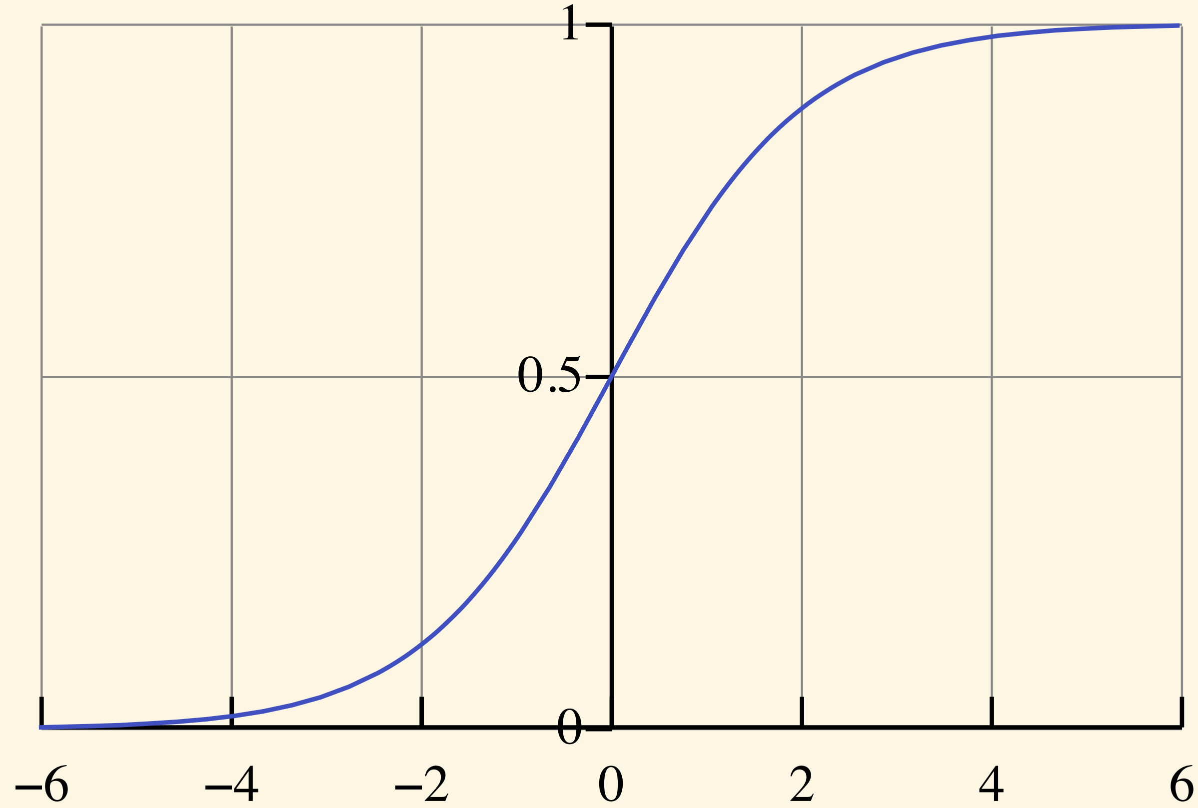
$$p_+ = e^{\mathbf{w}^T\mathbf{x}} - p_+ e^{\mathbf{w}^T\mathbf{x}}$$

$$p_+ + p_+ e^{\mathbf{w}^T\mathbf{x}} = e^{\mathbf{w}^T\mathbf{x}}$$

$$p_+(1 + e^{\mathbf{w}^T\mathbf{x}}) = e^{\mathbf{w}^T\mathbf{x}}$$

$$p_+ = \frac{e^{\mathbf{w}^T\mathbf{x}}}{1+e^{\mathbf{w}^T\mathbf{x}}}$$

$$p_+ = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}$$



WHAT'S THE PROBABILITY WHEN IT IS INTERESTING?

$$\mathrm{P}(G = k|X = x) = \frac{e^{\mathbf{w}_k^T \mathbf{x}}}{1+\sum_i^{K-1} e^{\mathbf{w}_i^T \mathbf{x}}}, k = 1, \ldots, K-1$$

$$\mathrm{P}(G = K|X = x) = \frac{1}{1+\sum_i^{K-1} e^{\mathbf{w}_i^T \mathbf{x}}}$$

SOFTMAX!

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

AN ALTERNATIVE PERSPECTIVE ON LOG ODDS

What's posterior probability of class c_1 given a sample $\mathbf{x}$?

$$\mathrm{p}(c_1|\mathbf{x}) = \frac{\mathrm{p}(\mathbf{x}|c_1)\mathrm{p}(c_1)}{\mathrm{p}(\mathbf{x}|c_1)\mathrm{p}(c_1) + \mathrm{p}(\mathbf{x}|c_2)\mathrm{p}(c_2)}$$

Let's introduce $a = \ln \frac{\mathrm{p}(\mathbf{x}|c_1)\mathrm{p}(c_1)}{\mathrm{p}(\mathbf{x}|c_2)\mathrm{p}(c_2)}$

$$\mathrm{p}(c_1|\mathbf{x}) = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

Nice properties of logistic sigmoid

$$\sigma(-a) = 1 - \sigma(-a)$$

$a = \ln\left(\frac{\sigma}{1-\sigma}\right)$ log odds???

$\frac{d\sigma}{da} = \sigma(1 - \sigma)$

MAXIMUM LIKELIHOOD ESTIMATE

$$l(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i}^{N} P_{\mathbf{w}}(c_k|x_i)$$

$$l(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i:\mathbf{x}_i \in c_1}^{N} P_{\mathbf{w}}(c_1|x_i) \prod_{i:\mathbf{x}_i \in c_2}^{N} P_{\mathbf{w}}(c_2|x_i)$$

$$l(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i:\mathbf{x}_i \in c_1}^{N} \sigma \prod_{i:\mathbf{x}_i \in c_2}^{N} (1 - \sigma)$$

$$l(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i}^{N} \sigma_i^{l_1} (1 - \sigma_i)^{1-l_1}$$

NEGATIVE LOG LIKELIHOOD

$$l(\mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i}^{N} \sigma_i^{l_1} (1 - \sigma_i)^{1-l_1}$$

$$\ell(\mathbf{w}) = -\sum_{i}^{N} (l_i \ln(\sigma_i) + (1 - l_i) \ln(1 - \sigma_i))$$

UNSUPERVISED AS SUPERVISED LEARNING

2009 - Second edition H.T.F.

Section 14.2.4

$x_1 \dots x_N \;\Big|\; x_1 \dots x_{N_0}$

$\frac{N_0}{N+N_0}$ $\frac{N}{N+N_0}$

$(g(x)+g_0(x))/2$

$\frac{g(x)}{g_0(x)}(y_i, x_i)$

$\mu(x) = P(Y|X) = \frac{g(x)}{g(x)+g_0(x)}$

$e^{f(x)}$

$\hat{g}(x) = g_0(x)\frac{\hat{\mu}(x)}{1-\hat{\mu}(x)}$

$= \prod g_0(x_i)$

NOISE CONTRASTIVE ESTIMATION

14.2.4

$$= \frac{1}{2T} \sum_t \ln[h(x_t;\theta)] + \ln[1 - h$$

$$h(u;\theta) = \frac{1}{1 + e^{(-G(u;\theta))}}$$

$$G(u;\theta) = \ln p_m(u;\theta) - $$

$$\ln \frac{p_m(u;}{p_n(u}$$

WORD2VEC AND NEGATIVE SAMPLING

NEGATIVE SAMPLING IN SKIP GRAM W2V

Mikolov's et al. version