Generalized Linear Models

Oswin Krause, PML, 2021

UNIVERSITY OF COPENHAGEN

Contents

We focus on

- Describing random variables as transformations of other random variables
- Using transformations to model Datasets and tasks
- Applying Bayesian and Frequentist methods to learn these transformations

Reminder: The normal distribution

- "X has a normal distribution with mean μ and variance $\sigma^2 > 0$":

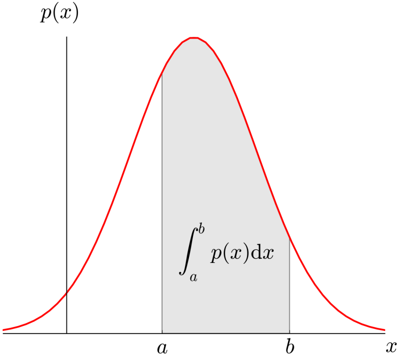
$$X \sim \mathcal{N}(\mu, \sigma^2)$$

- Probability Density Function (pdf)

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Cumulative Density function (cdf)

$$P(X \leq b) = \int_{-\infty}^{b} p(x)dx$$



$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$

Reminder: Properties of the normal distribution

Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, $a, b \in \mathbb{R}$, $b \neq 0$
Then, we have the following properties:

- Affine transformations: $Z = a + bX$ is normal distributed and

$$Z \sim \mathcal{N}(a + b\mu_X, b^2\sigma_X^2)$$

$\Rightarrow$ $X = \mu_X + \sigma_X\epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$

- Summation: $Z = X + Y$ is normal distributed and

$$Z \sim \mathcal{N}(\mu_X + \mu_y, \sigma_X^2 + \sigma_Y^2)$$

$\Rightarrow$ All linear combinations of normal random variables are normal random variables.

Definition: Multivariate normal distribution

Let $\epsilon \in \mathbb{R}^N$ be a random variable with elements distributed as $\epsilon_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, N$. Further, let $A \in \mathbb{R}^{d \times N}$, $\mu \in \mathbb{R}^d$. A random variable of the form

$$X = \mu + A\epsilon$$

is called Multivariate Normal Distributed with mean $E[X] = \mu$ and variance $\Sigma = AA^T$, or in short

$$X \sim \mathcal{N}(\mu, \Sigma) \quad .$$

Usual Definition: Multivariate normal distribution

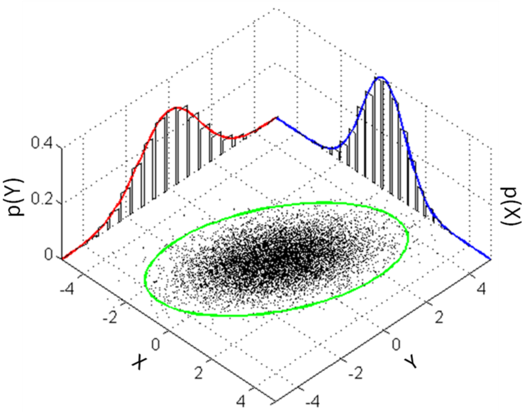
Let $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ symmetric positive definite.

- We say $X \sim \mathcal{N}(\mu, \Sigma)$ if it has pdf

$$p(x) = \mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

- Not all multivariate normal distributions have a pdf. We will see this next week.

Usual Definition: Multivariate normal distribution



MVN: Closed under Linear transformations

Let $X \sim \mathcal{N}(\mu_x, \Sigma_X)$ be a d-dimensional multivariate random variable. Further, let $Q \in \mathbb{R}^{k \times d}$, then

$$Z = QX$$

is a multivariate normal random variable with $Z \sim \mathcal{N}(Q\mu_X, Q\Sigma_X Q^T)$

Proof

Let $\epsilon \sim \mathcal{N}(0, I_d)$. Further, let A_X be a matrix such, that $\Sigma_X = A_X A_X^T$. Then $X = \mu_X + A_X \epsilon$ and

$$Z = QX = Q(\mu_X + A_X \epsilon) = \underbrace{Q\mu_X}_{\mu_Z} + \underbrace{QA_X}_{A_Z} \epsilon = \mu_Z + A_Z \epsilon.$$

This meets the definition of a multivariate normal distribution and the Covariance Matrix is

$$\Sigma_Z = A_Z A_Z^T = QA_X A_X^T Q^T = Q\Sigma_X Q^T \quad .$$

Reminder: Marginal and Conditional distribution of Multivariate Normal

Let $X \sim \mathcal{N}(\mu, \Sigma)$
Partition vector and matrix into blocks of size K and $N-K$

$$X = \left[\begin{array}{c} X_1 \in \mathbb{R}^K \\ \hline X_2 \in \mathbb{R}^{N-K} \end{array}\right], \ \mu = \left[\begin{array}{c} \mu_1 \\ \hline \mu_2 \end{array}\right], \ \Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$$

- Marginal distribution

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$$

- X_2 conditioned on X_1,

$$X_2 | X_1 \sim \mathcal{N}(\mu_{2|1}, \Sigma_{2|1})$$

where

$$\mu_{2|1} = \mu_2 + \Sigma_{12}\Sigma_{11}^{-1}(X_1 - \mu_1), \qquad \Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{21}^T$$

Reminder: Joint from Conditional

Let $X \in \mathbb{R}^N \sim \mathcal{N}(\mu_X, \Sigma_X)$
Let $Y|X \sim \mathcal{N}(\mu_Y + AX, \Sigma_Y)$
Then, the joint distribution of X and Y is:

$$\left[\begin{array}{c} X \\ \hline Y \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \mu_X \\ \hline \mu_Y + A\mu_X \end{array}\right], \left[\begin{array}{c|c} \Sigma_X & \Sigma_X A^T \\ \hline A\Sigma_X & \Sigma_Y + A\Sigma_X A^T \end{array}\right]\right)$$

Trick: Bring in Linear transformation form

Let $X \sim \mathcal{N}(\mu_X, \Sigma_X)$
Write as linear transformation with $\Sigma_X = A_X A_X^T$

$$X = \mu_X + A_X \epsilon_X, \ \epsilon_X \sim \mathcal{N}(0, I)$$

Trick: Bring in Linear transformation form

Let $X \sim \mathcal{N}(\mu_X, \Sigma_X)$
Write as linear transformation with $\Sigma_X = A_X A_X^T$

$$X = \mu_X + A_X \epsilon_X, \ \epsilon_X \sim \mathcal{N}(0, I)$$

We can do the same for $Y \sim \mathcal{N}(\mu_Y + AX, \Sigma_Y)$ and $\Sigma_X = A_Y A_Y^T$

$$Y = \mu_Y + AX + A_y \epsilon_Y, \ \epsilon_Y \sim \mathcal{N}(0, I)$$

Trick: Bring in Linear transformation form

Let $X \sim \mathcal{N}(\mu_X, \Sigma_X)$
Write as linear transformation with $\Sigma_X = A_X A_X^T$

$$X = \mu_X + A_X \epsilon_X,\ \epsilon_X \sim \mathcal{N}(0, I)$$

We can do the same for $Y \sim \mathcal{N}(\mu_Y + AX, \Sigma_Y)$ and $\Sigma_X = A_Y A_Y^T$

$$Y = \mu_Y + AX + A_y \epsilon_Y,\ \epsilon_Y \sim \mathcal{N}(0, I)$$

Insert X in Y

$$\begin{aligned} Y &= \mu_y + A(\mu_X + A_X \epsilon_X) + A_y \epsilon_Y \\ &= \mu_Y + A\mu_X + AA_X \epsilon_X + A_y \epsilon_Y \end{aligned}$$

Trick: Bring in Linear transformation form

We have

$$
\begin{aligned}
X &= \mu_X + A_X \epsilon_X, \; \epsilon_X \sim \mathcal{N}(0, I) \\
Y &= \mu_Y + A\mu_X + AA_X \epsilon_X + A_y \epsilon_Y, \; \epsilon_Y \sim \mathcal{N}(0, I)
\end{aligned}
$$

Write in Block Matrix form

Trick: Bring in Linear transformation form

We have

$$
\begin{aligned}
X &= \mu_X + A_X \epsilon_X, \ \epsilon_X \sim \mathcal{N}(0, I) \\
Y &= \mu_Y + A\mu_X + AA_X \epsilon_X + A_y \epsilon_Y, \ \epsilon_Y \sim \mathcal{N}(0, I)
\end{aligned}
$$

Write in Block Matrix form

$$
\left[\begin{array}{c} X \\ \hline Y \end{array}\right] = \left[\begin{array}{c} \mu_X \\ \hline \mu_Y + A\mu_X \end{array}\right] + \left[\begin{array}{c|c} A_X & 0 \\ \hline AA_X & A_Y \end{array}\right] \left[\begin{array}{c} \epsilon_X \\ \hline \epsilon_Y \end{array}\right]
$$

Trick: Bring in Linear transformation form

We have

$$X = \mu_X + A_X \epsilon_X, \ \epsilon_X \sim \mathcal{N}(0, I)$$
$$Y = \mu_Y + A\mu_X + AA_X \epsilon_X + A_y \epsilon_Y, \ \epsilon_Y \sim \mathcal{N}(0, I)$$

Write in Block Matrix form

$$\underbrace{\left[\begin{array}{c} X \\ \hline Y \end{array}\right]}_{Z} = \underbrace{\left[\begin{array}{c} \mu_X \\ \hline \mu_Y \end{array}\right]}_{\mu_Z} + \underbrace{\left[\begin{array}{c|c} A_X & 0 \\ \hline A & A_Y \end{array}\right]}_{A_Z} \underbrace{\left[\begin{array}{c} \epsilon_X \\ \hline \epsilon_Y \end{array}\right]}_{\epsilon}$$

Thus, $Z \sim \mathcal{N}(\mu_Z, \Sigma_Z)$, $\Sigma_Z = A_Z A_Z^T$. A_Z is invertible, since A_X and A_Y are and the upper and lower block are linearly independent.

Final: Construct Σ_Z

We have

$$\Sigma_Z = \left[\begin{array}{c|c} A_X & 0 \\ \hline AA_X & A_Y \end{array}\right] \left[\begin{array}{c|c} A_X^T & A_X^T A^T \\ \hline 0 & A_Y^T \end{array}\right]$$

Final: Construct Σ_Z

We have

$$\Sigma_Z = \left[\begin{array}{c|c} A_X & 0 \\ \hline AA_X & A_Y \end{array}\right] \left[\begin{array}{c|c} A_X^T & A_X^T A^T \\ \hline 0 & A_Y^T \end{array}\right]$$

$$= \left[\begin{array}{c|c} A_X A_X^T & A_X A_X^T A^T \\ \hline AA_X A_X^T & A_Y^T A_Y \end{array}\right]$$

Final: Construct Σ_Z

We have

$$\Sigma_Z = \left[\begin{array}{c|c} A_X & 0 \\ \hline AA_X & A_Y \end{array}\right] \left[\begin{array}{c|c} A_X^T & A_X^T A^T \\ \hline 0 & A_Y^T \end{array}\right]$$

$$= \left[\begin{array}{c|c} A_X A_X^T & A_X A_X^T A^T \\ \hline AA_X A_X^T & A_Y^T A_Y \end{array}\right]$$

Insert $\Sigma_X = A_X A_X^T$, $\Sigma_Y = A_Y A_Y^T$

$$\Sigma_Z = \left[\begin{array}{c|c} \Sigma_X & \Sigma_X A^T \\ \hline A\Sigma_X & \Sigma_Y \end{array}\right]$$

Probabilistic modeling

Probabilistic models

- Model an event or phenomenon by a probability distribution
- Different sources of randomness
  - Imprecision in measurement (Noise)
  - Missing observations
  - Stochasticity inherent to a process (predicting the future...)
- Applications
  - Estimate expected costs or risks
  - Estimate unknown variables based on the observed values
  - Learn relationship between variables

Probabilistic Model: Regression

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \mathbb{R}$, find relationship $y = g(x)$

Data is generated as:

1. Randomly sample point $x \sim p(x)$

Probabilistic Model: Regression

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \mathbb{R}$, find relationship $y = g(x)$
Data is generated as:

1. Randomly sample point $x \sim p(x)$
2. True label $y_{\text{true}} = g(x)$ can not be observed

Probabilistic Model: Regression

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \mathbb{R}$, find relationship $y = g(x)$
Data is generated as:

1. Randomly sample point $x \sim p(x)$
2. True label $y_{\text{true}} = g(x)$ can not be observed
3. We observe corrupted label $y = g(x) + \epsilon$
   ϵ is sample of noise distribution

Probabilistic Model: Regression

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \mathbb{R}$, find relationship $y = g(x)$
Data is generated as:

1. Randomly sample point $x \sim p(x)$
2. True label $y_{\text{true}} = g(x)$ can not be observed
3. We observe corrupted label $y = g(x) + \epsilon$
   ϵ is sample of noise distribution

Learning the model: Given dataset $\mathcal{D}$ find $f \approx g$.

Example: Model of Linear regression

We assume

- g is linear combination of basis functions

Example: Model of Linear regression

We assume

- g is linear combination of basis functions
- $g(x) = f_\theta(x) = \theta^T \phi(x)$ for some $\theta \in \mathbb{R}^k$ and predefined $\phi(x) : \mathbb{R}^d \to \mathbb{R}^k$
- For example $\phi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$

Example: Model of Linear regression

We assume

- g is linear combination of basis functions
- $g(x) = f_\theta(x) = \theta^T \phi(x)$ for some $\theta \in \mathbb{R}^k$ and predefined $\phi(x) : \mathbb{R}^d \to \mathbb{R}^k$
- For example $\phi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$
- Label noise distribution $\epsilon \sim \mathcal{N}(0, \sigma_Y^2)$

Example: Model of Linear regression

We assume

- g is linear combination of basis functions
- $g(x) = f_\theta(x) = \theta^T\phi(x)$ for some $\theta \in \mathbb{R}^k$ and predefined $\phi(x) : \mathbb{R}^d \to \mathbb{R}^k$
- For example $\phi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$
- Label noise distribution $\epsilon \sim \mathcal{N}(0, \sigma_Y^2)$
- Label distribution $p(y|x,\theta) = \mathcal{N}(y; \theta^T\phi(x), \sigma_Y^2)$
  Proof: $y = g(x) + \epsilon = \theta^T\phi(x) + \epsilon$ is an affine transformation of ϵ

Example: Model of Linear regression

We assume

- g is linear combination of basis functions
- $g(x) = f_\theta(x) = \theta^T\phi(x)$ for some $\theta \in \mathbb{R}^k$ and predefined $\phi(x) : \mathbb{R}^d \to \mathbb{R}^k$
- For example $\phi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$
- Label noise distribution $\epsilon \sim \mathcal{N}(0, \sigma_Y^2)$
- Label distribution $p(y|x,\theta) = \mathcal{N}(y; \theta^T\phi(x), \sigma_Y^2)$
  Proof: $y = g(x) + \epsilon = \theta^T\phi(x) + \epsilon$ is an affine transformation of ϵ

Learning the model: Given dataset $\mathcal{D}$ find θ.

Implementation: Bayesian Linear Regression

- Data:$\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots (x^{(\ell)}, y^{(\ell)})\}$
- $p(y^{(i)}|x^{(i)}, \theta) = \mathcal{N}(y^{(i)}; \theta^T \phi(x^{(i)}), \sigma_y^2)$
- New: Prior $\theta \sim \mathcal{N}(0, I)$

We need to find

$$p(\theta|\mathcal{D}) = \frac{p(\theta) \prod_{i=1}^{\ell} p(y^{(i)}|x^{(i)}, \theta)}{p(\mathcal{D})}$$

Idea: First compute $p(\theta, y_1, \ldots, y_n|x_1, \ldots, x_n)$, then condition on y_i

Implementation: Bayesian Linear Regression

Let $\Phi \in \mathbb{R}^{\ell \times k}$ a matrix with $\phi(x^{(i)})$ being the ith row and $y \in \mathbb{R}^{\ell}$ the vector of y_i.
We have

$$p(\theta, y|\Phi) = p(\theta) \prod_{i=1}^{\ell} \mathcal{N}(y^{(i)}; \theta^T \phi(x^{(i)}), \sigma_y^2) = p(\theta)\mathcal{N}(y; \Phi\theta, \sigma_y^2 I)$$

We use the rule of the joint distribution

Implementation: Bayesian Linear Regression

Let $\Phi \in \mathbb{R}^{\ell \times k}$ a matrix with $\phi(x^{(i)})$ being the ith row and $y \in \mathbb{R}^{\ell}$ the vector of y_i. We have

$$p(\theta, y|\Phi) = p(\theta)\prod_{i=1}^{\ell}\mathcal{N}(y^{(i)}; \theta^T\phi(x^{(i)}), \sigma_y^2) = p(\theta)\mathcal{N}(y; \Phi\theta, \sigma_y^2 I)$$

We use the rule of the joint distribution

$$p(\theta, y|\Phi) = \mathcal{N}\left(\left[\begin{array}{c}\theta \\ \hline y\end{array}\right]; \left[\begin{array}{c}0 \\ \hline 0\end{array}\right], \left[\begin{array}{c|c} I & \Phi^T \\ \hline \Phi & \sigma_y^2 I + \Phi\Phi^T\end{array}\right]\right)$$

Implementation: Bayesian Linear Regression

We have (permuted θ and y)

$$p(\theta, y|\Phi) = \mathcal{N}\left(\left[\begin{array}{c} y \\ \hline \theta \end{array}\right]; \left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right], \left[\begin{array}{c|c} \sigma_y^2 I + \Phi\Phi^T & \Phi \\ \hline \Phi^T & I \end{array}\right]\right)$$

Now we use the conditional rule on y, leading to

$$p(\theta|\mathcal{D}) = p(\theta|y, \Phi) = \mathcal{N}(\theta; \mu_{\theta|\mathcal{D}}, \Sigma_{\theta|\mathcal{D}})$$

with

$$\begin{aligned} \mu_{\theta|\mathcal{D}} &= \Phi^T(\sigma_y^2 I_\ell + \Phi\Phi^T)^{-1} y \\ \Sigma_{\theta|\mathcal{D}} &= I - \Phi^T(\sigma_y^2 I_\ell + \Phi\Phi^T)^{-1}\Phi \quad . \end{aligned}$$

Implementation: Bayesian Linear Regression

Posterior predictive: distribution of labels $\hat{y}$ for query point x

$$p(\hat{y}|x, \mathcal{D}) = \int p(\theta|\mathcal{D}) p(\hat{y}|x, \theta)\, d\theta$$

Same trick: first compute joint distribution $p(y, \theta|\mathcal{D})$, then marginalize θ.

$$p(\hat{y}|x, \mathcal{D}) = \mathcal{N}(\hat{y}; x^T \mu_{\theta|\mathcal{D}}, \sigma_y^2 + x^T \Sigma_{\theta|\mathcal{D}} x)$$

Binary Classification: Generative Model

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \{0, 1\}$, find relationship $y = h(x)$

1. Randomly sample point $x \sim p(x)$

Binary Classification: Generative Model

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \{0, 1\}$, find relationship $y = h(x)$

1. Randomly sample point $x \sim p(x)$
2. True label

$$y_{\text{true}} = h(x) = \begin{cases} 1, & \text{if } g(x) > 0 \\ 0, & \text{otherwise} \end{cases} .$$

Binary Classification: Generative Model

Goal: Given dataset $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots (x^{(\ell)}, y^{(\ell)})\}$, $y \in \{0, 1\}$, find relationship $y = h(x)$

1. Randomly sample point $x \sim p(x)$
2. True label

$$y_{\text{true}} = h(x) = \begin{cases} 1, & \text{if } g(x) > 0 \\ 0, & \text{otherwise} \end{cases} .$$

3. We observe corrupted label

$$y = \begin{cases} 1, & \text{if } g(x) + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} .$$

ϵ is sample of noise distribution

Distribution of corrupted labels

We have:

$$y = \begin{cases} 1, & \text{if } g(x) + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} .$$

What is $p(y = 1|x)$?

Distribution of corrupted labels

We have:

$$y = \begin{cases} 1, & \text{if } g(x) + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} .$$

What is $p(y = 1|x)$?

$$P(y = 1|x) = P(\epsilon > -g(x))$$

Distribution of corrupted labels

We have:

$$y = \begin{cases} 1, & \text{if } g(x) + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} .$$

What is $p(y = 1|x)$?

$$\begin{aligned} P(y = 1|x) &= P(\epsilon > -g(x)) \\ &= 1 - P(\epsilon \leq -g(x)) \end{aligned}$$

Distribution of corrupted labels

We have:

$$y = \begin{cases} 1, & \text{if } g(x) + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases} \quad .$$

What is $p(y = 1|x)$?

$$\begin{aligned} P(y = 1|x) &= P(\epsilon > -g(x)) \\ &= 1 - P(\epsilon \leq -g(x)) \\ &= 1 - \int_{-\infty}^{-g(x)} p(\epsilon)\, d\epsilon \quad . \end{aligned}$$

$P(\epsilon \leq t)$ is the cumulative distribution function of ϵ

Linear Probit regression

- g is approximately linear combination of basis functions
- $g(x) \approx f_\theta(x) = \theta^T \phi(x)$ for some θ and predefined $\phi(x)$

Linear Probit regression

- g is approximately linear combination of basis functions
- $g(x) \approx f_\theta(x) = \theta^T \phi(x)$ for some θ and predefined $\phi(x)$
- Noise distribution $\epsilon \sim \mathcal{N}(0, 1)$

Linear Probit regression

- g is approximately linear combination of basis functions
- $g(x) \approx f_\theta(x) = \theta^T \phi(x)$ for some θ and predefined $\phi(x)$
- Noise distribution $\epsilon \sim \mathcal{N}(0, 1)$
- $p(y = 1|x, \theta) = 1 - P(\epsilon \leq -\theta^T \phi(x)) = P(\epsilon \leq \theta^T \phi(x))$
  (No closed form solution)

Linear Logistic regression

- $g(x) \approx f_\theta(x) = \theta^T \phi(x)$
- $\epsilon \sim \text{Logistic}(0, 1)$

$$p(\epsilon) = \frac{\exp(-\epsilon)}{(1 + \exp(-\epsilon))^2} \quad .$$

- $p(y = 1|x, \theta) = \frac{1}{1+\exp(-\theta^T x))}$
- This is what we used earlier!