Logistic Regression

Input: d-dimension feature vector $x \in \mathbb{R}^d$
Output: a class or label $l \in \{1, \ldots, k\}$

Ex: Image classification:
Given input image $x \in \mathbb{R}^{n \times n}$, predict a label for the image (e.g. cat/dog, MNIST: $\{0, \ldots, 9\}$, CIFAR-10 $\{0, \ldots, 9\}$)

Ex: Binary Classification:
Given some medical data $\leftarrow x \in \mathbb{R}^n$
try to predict incidence of heart disease
label: $\{0/1\}$

Logistic Regression model

Given input feature vector $x \in \mathbb{R}^n$,
the model predicts a prob. distribution over the labels $\{1, \ldots, k\}$

$$\text{softmax}\underbrace{\left(\underbrace{A}_{\in \mathbb{R}^{k \times n}} \underbrace{x}_{\in \mathbb{R}^n} + b\right)}_{\in \mathbb{R}^k}$$

affine linear map: $\mathbb{R}^n \to \mathbb{R}^k \xrightarrow{\text{softmax}} P(\{1, \ldots, k\})$

softmax: Input: $y \in \mathbb{R}^k = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$

Output: Distribution $p(j) = \frac{e^{y_j}}{\sum_{i=1}^{k} e^{y_i}}$

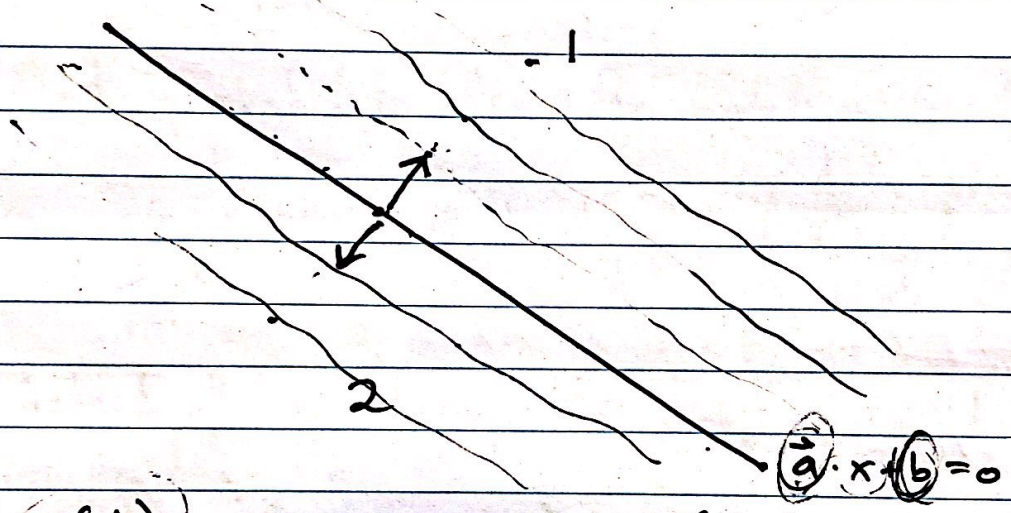
"Take exponential and normalize"

Given input $x \in \mathbb{R}^n$, we return the distribution

$\text{softmax}(Ax+b)$, A, b are parameters.

Ex: Two classes: Parameters are $\vec{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$.

$$p(1) = \frac{e^{\vec{a}\cdot x + b}}{e^{\vec{a}\cdot x + b} + 1}, \quad p(2) = \frac{1}{e^{\vec{a}x+b} + 1}$$



$$\frac{p(1)}{p(2)} = e^{\vec{a}\cdot x + b}, \quad \log\left(\frac{p(1)}{p(2)}\right) \leftarrow \text{logarithm of the odds}$$

Assumption: $\log\left(\frac{p(1)}{p(2)}\right)$ is linear in the feature vector.

Learning the parameters $\vec{a}, b$ from data

Data: $\{(x_1, l_1), \ldots, (x_n, l_n)\} = D$

$x_1, \ldots, x_n$: Feature vectors; $l_1, \ldots, l_n$: labels

Given this data D, how can we estimate A, b?

Maximum Likelihood

Given x_1, model gives $\text{softmax}(Ax_1 + b)$

$$\frac{1}{\sum_{i=1}^{k} e^{a_i \cdot x_1 + b_i}} \begin{pmatrix} e^{\vec{a}_1 \cdot x_1 + b_1} \\ \vdots \\ e^{\vec{a}_k \cdot x_1 + b_k} \end{pmatrix}_k$$

What probability does the model asrign to ℓ_1?

$$\frac{1}{\sum_{i=1}^{k} e^{a_i \cdot x_1 + b_i}} \cdot e^{\vec{a}_{\ell_1} \cdot x_1 + b_{\ell_1}}$$

(data points are indep. so take product)

Instead of considering this prob, we consider its negative logarithm:

$$\log\left(\sum_{i=1}^{k} e^{a_i \cdot x_1 + b_i}\right) - \left(\vec{a}_{\ell_1} \cdot x_1 + b_{\ell_1}\right)$$

Logistic Regression loss:

$$L_D(A,b) = \sum_{(x,\ell) \in D} \left[\log\left(\sum_{i=1}^{k} a_i \cdot x + b_i\right) - \left(\vec{a}_\ell \cdot x + b_\ell\right)\right]$$

$$(A,b) = \min_{A,b} L_D(A,b)$$

Basic Statistical Learning Theory

- Goal: Estimate an unknown probability dist. D on a set X from samples (i.i.d) $x_1, \ldots, x_n \in X$.

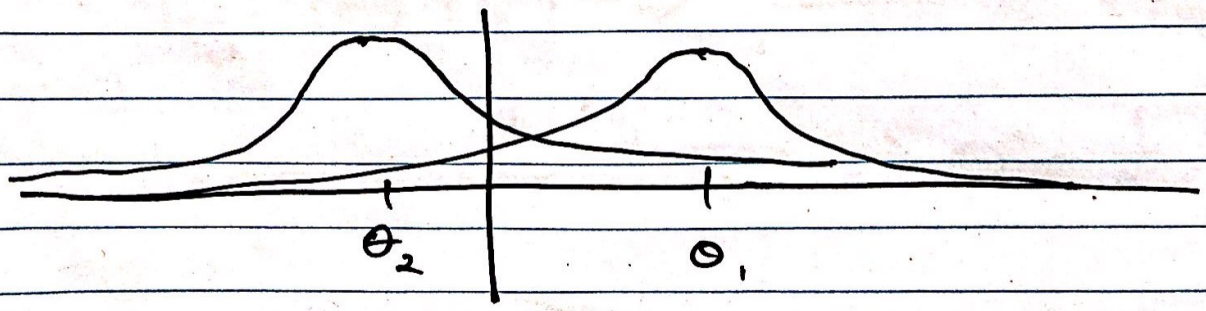
- Introduce a family of distributions p_θ for $\theta \in H$ and try to choose θ to "match" the samples.

- Maximum Likelihood Estimate: Choose θ to maximize the probability of the samples.

Ex: Let $X = \mathbb{R}$, have some samples $x_1, \ldots, x_n$ drawn from a distribution D.

- Let p_θ be a Gaussian w/ variance 1, centered at $\theta \in \mathbb{R} = H$, i.e.

density $p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}$ (centered at θ; variance 1)



- Use the samples to find the center θ.

Maximum Likelihood Estimation

- Given $\theta \in \Theta = \mathbb{R}$, what is the probability of $\{x_j\}_{j=1}^n$?

Likelihood function (as a function of θ)

- Samples independent: $p_\theta(\{x_j\}_{j=1}^n) = \prod_{j=1}^n p_\theta(x_j)$.

$$= \frac{1}{(2\pi)^{n/2}} \prod_{j=1}^n e^{-(x_j-\theta)^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{j=1}^n (x_j-\theta)^2/2}.$$

- MLE: Choose θ to maximize this!
- Often it's useful to consider $\log(p_\theta(\{x_j\}_{j=1}^n))$.

log likelihood function

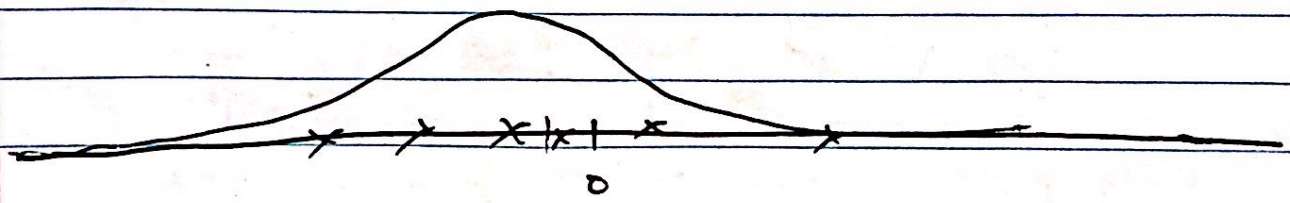
$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} \log(p_\theta(\{x_j\}_{j=1}^n))$$

$$\left(\operatorname{argmin}_{\theta \in \Theta} -\log(p_\theta(\{x_j\}_{j=1}^n))\right)$$

- For this example: $\log(p_\theta(\{x_j\}_{j=1}^n)) =$

$$= -\log(2\pi)\cdot\left(\frac{n}{2}\right) - \sum_{j=1}^n \frac{(x_j-\theta)^2}{2}.$$

$$\theta^* = \operatorname{argmin}_{\theta \in \mathbb{R}} \sum_{j=1}^n \frac{(x_j-\theta)^2}{2} \Rightarrow \theta^* = \frac{1}{n}\sum_{j=1}^n x_j.$$



Classification / Logistic Regression

- For Classification $\tilde{X} = X \times Y$
  - X: features, Y: labels
- We have samples $\{(x_j, y_j)\}_{j=1}^n$.
- Suppose that $\mathcal{D}$ is an unknown distribution on $\tilde{X}$, but we're only trying to estimate $\mathcal{D}_{Y|X}$ as a function of x.
- Introduce parameters $\Theta \in \Theta$, we define $p(y|x,\Theta)$ ← called the model
- Now choose Θ to model the data $\{(x_j, y_j)\}_{j=1}^n$.
- MLE:

$$\Theta^* = \underset{\Theta\in\Theta}{\operatorname{argmax}}\, p\left(\{y_j\}_{j=1}^n \,\middle|\, \{x_j\}_{j=1}^n, \Theta\right) =$$

$$= \underset{\Theta\in\Theta}{\operatorname{argmax}} \prod_{j=1}^n p(y_j|x_j,\Theta)$$

$$= \underset{\Theta\in\Theta}{\operatorname{argmin}} \sum_{j=1}^n -\log\left(p(y_j|x_j,\Theta)\right).$$

- Ex: Logistic Regression: $W \in \mathbb{R}^{k\times d}$, $b \in \mathbb{R}^k$.
  - Model:

$$p(y|x,\Theta) = p(y|x,W,b)$$

  - d - dimension of features, i.e. number of pixels
  - k - number of classes.

$W \in \mathbb{R}^{k \times d}$, $x \in \mathbb{R}^d$, $b \in \mathbb{R}^k$

$$p(y \mid x, W, b) = \text{softmax}\left(\underbrace{Wx + b}_{\in \mathbb{R}^k}\right) =$$

$$= \frac{1}{\sum_{i=1}^{k} e^{w_i \cdot x + b_i}} \begin{pmatrix} e^{w_1 x + b_1} \\ \vdots \\ e^{w_k x + b_k} \end{pmatrix}$$

$$y_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \leftarrow \text{label } i \\ 0 \\ 0 \end{pmatrix} = e_{y_i}$$

- Need to calculate:

$$-\log\left(p(y_i \mid x_i, W, b)\right) = -\log\left(\frac{1}{e^{Wx_i + b} \cdot \mathbb{1}} e^{Wx_i + b} \cdot y_i\right)$$

$$= \log\left(e^{Wx_i + b} \cdot \mathbb{1}\right) - \log\left(e^{Wx_i + b} \cdot y_i\right)$$

$$(W, b)^* = \underset{W, b}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(e^{Wx_i + b} \cdot \mathbb{1}\right) - \log\left(e^{Wx_i + b} \cdot y_i\right)$$

$$\in \mathbb{R}^{k \times d} \times \mathbb{R}^k$$

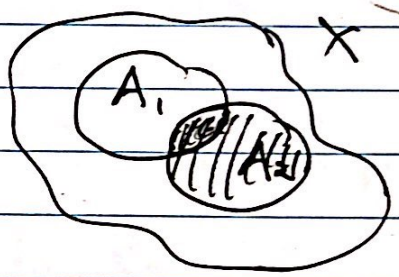
Bayesian Approach to Machine Learning

- Goal: Estimate an unknown distribution on X from data $\{x_j\}_{j=1}^{n}$.
  - Build a model:
    - Set of parameters (H)
    - Family of distributions on X, $p(x|\theta)$ ← parameter $\in (H)$
    - Prior distribution on the parameters (H). $q(\theta)$.
  - Use Bayes' Law:

$$p(\theta|x)\,p(x) = p(x \text{ and } \theta) = p(x|\theta)\,p(\theta)$$

  - Recall if A_1, A_2 events:

$$p(A_1|A_2) = \frac{p(A_1 \cap A_2)}{p(A_2)}$$



$$p(\theta|x) = \frac{p(x|\theta)\,q(\theta)}{p(x)}$$

$$p(\theta|x) \sim p(x|\theta)\,q(\theta)$$

posterior distribution ← $p(\theta|x)$; likelihood function ← $p(x|\theta)$; prior distribution ← $q(\theta)$

- Start with: prior $q(\theta)$
- Add data x, replace q with the posterior $p(\theta|x) \sim p(x|\theta)\,q(\theta)$
- More data → multiply by Likelihood + normalize.

- Left with a posterior distribution
  - Sample from posterior dist to approximate

$$P_{pred}(x) = \int_{\Theta} p(x|\theta)\, p(\theta)\, d\theta$$

($p(\theta)$: posterior dist)

- Choose θ to maximize posterior distribution:

$$\theta^* = \operatorname{argmax}_{\theta \in \Theta} p(\theta|x).$$

$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} -\log(p(\theta|x)) =$$

$$= \operatorname{argmin}_{\theta \in \Theta} -\log(p(x|\theta)) - \log(q(\theta)) + \log(p(x))$$

$$= \operatorname{argmin}_{\theta \in \Theta} \underbrace{-\log(p(x|\theta))}_{\text{Negative log likelihood}} \underbrace{-\log(q(\theta))}_{\text{Regularization coming from prior.}}$$

Ex: Image Classification / Logistic Regression

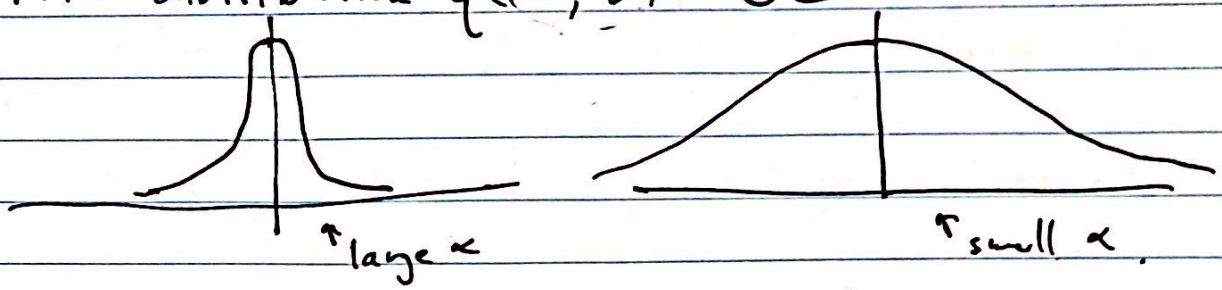
- Images $x \in X = \mathbb{R}^d$
- Labels $y \in Y = \{e_1, \dots, e_k\}$, k = # of labels

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix}$$

- Data: $\{(x_j, y_j)\}_{j=1}^{n}$.
- Model: $\theta = (W, b) \in \mathbb{R}^{k \times d} \times \mathbb{R}^k$.

$$p(y|x, \theta) = \frac{e^{Wx+b} \cdot y}{e^{Wx+b} \cdot \mathbb{1}}.$$

- Prior distribution $q(W, b) = Ce^{-\alpha(\|W\|_2^2 + \|b\|_2^2)}$



↑ large α ↑ small α.

- Calculate the posterior:

$$p(W, b \mid x, y) = \frac{p(y \mid x, W, b)\, q(W, b)}{p(y \mid x)}$$

$$\hookrightarrow = \int_{(H)} p(y \mid x, W, b)\, q(W, b)\, d\theta.$$

$$(W, b)^* = \underset{W,b}{\operatorname{argmin}} -\log\left(p(W, b \mid \{x_j, y_j\}_{j=1}^n)\right) =$$

$$= \underset{W,b}{\operatorname{argmin}} -\log\left(p(\{y_j\}_{j=1}^n \mid \{x_j\}_{j=1}^n, W, b)\right) - \log\left(q(W, b)\right)$$

$$= \underset{W,b}{\operatorname{argmin}} -\log\left(\prod_{j=1}^n \underbrace{p(y_j \mid x_j, W, b)}_{\frac{e^{Wx_j+b} \cdot y_j}{e^{Wx_j+b} \cdot 1\!\!1}}\right) - \log\left(\underbrace{q(W, b)}_{Ce^{-\alpha(\|W\|_2^2 + \|b\|_2^2)}}\right)$$

$$= \underset{W,b}{\operatorname{argmin}} \sum_{j=1}^n \log\left(e^{Wx_j+b} \cdot 1\!\!1\right) - \log\left(e^{Wx_j+b} \cdot y_j\right) + \alpha\left(\|W\|_2^2 + \|b\|_2^2\right).$$