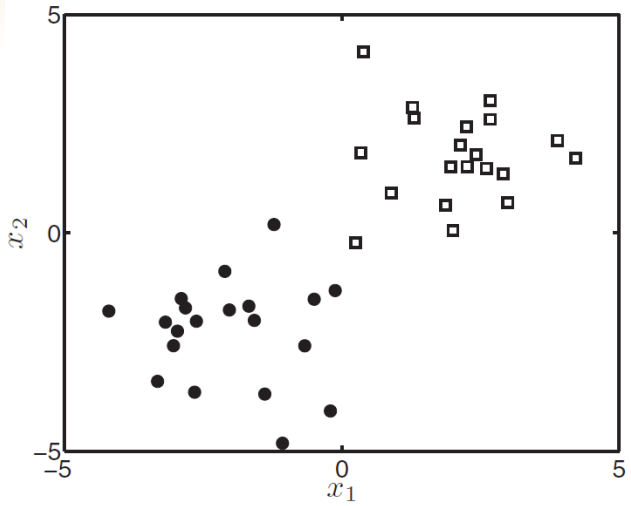
ECE 469/ECE 568 - Machine Learning

September 30, 2024

Bayesian Inference/Learning and Classification

- Within the Bayesian learning framework, all unknown quantities are typically treated as random variables.
- Thus, each unknown parameter is described by a probability distribution rather than an individual value.
- Consequently, the uncertainty in the parameter estimates is accounted into future predictions.

Binary responses/targets



- Each object is described by two attributes denoted by x_1 and x_2.
- Each object also has a binary response/target, $y \in \{0, 1\}$.
- The objects can be plotted with a symbol that depends on their response/target.
- For example, if $y = 0$, the object is plotted as a circle, where as if $y = 1$ is a square.
- Our objective is to use this data to build a model that will enable us to predict/classify the response/targets.
- This task is known as classification, which is one of the major problems within machine learning.

Bayesian model for binary responses/targets

- Our data can be represented via vectors and matrices as

$$\mathbf{x}^{(i)} = [x_1^{(i)} x_2^{(i)}]^T, \qquad \mathbf{w} = [w_1 w_2]^T, \qquad \mathbf{X} = [(\mathbf{x}^{(1)})^T, \cdots, (\mathbf{x}^{(N)})^T]^T$$

- The posterior density over the parameters of the model is given by using Bayes' rule as

$$f(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{f(\mathbf{y}, \mathbf{w}|\mathbf{X})}{f(\mathbf{y}|\mathbf{X})} = \frac{f(\mathbf{y}|\mathbf{X}, \mathbf{w}) f(\mathbf{w})}{f(\mathbf{y}|\mathbf{X})}$$

- Here, the marginal likelihood function is given by

$$f(\mathbf{y}|\mathbf{X}) = \int f(\mathbf{y}|\mathbf{X}, \mathbf{w}) f(\mathbf{w}) d\mathbf{w}$$

Bayesian model for binary responses/targets

- The prior distribution of $\mathbf{w}$ can be assumed to Gaussian distributed $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.
- The likelihood function is given by

$$f(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{N} f(y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}),$$

  where $y^{(i)}$ is a binary variable indicating the class $y^{(i)} \in \{1, 0\}$.
- Thus, $y^{(i)}$ must be modeled as a binary random variable ($Y^{(i)}$).

$$f(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{N} P(Y^{(i)} = y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}),$$

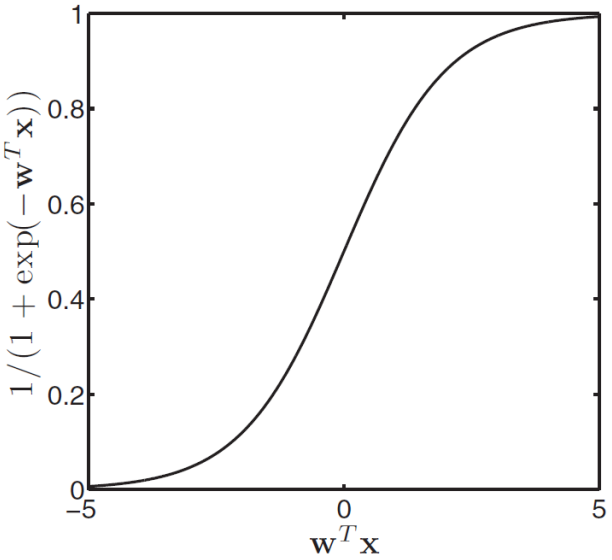
- This likelihood function will be high if the model assigns high probabilities for class 1 when we observe class 1
- It will also high probabilities for class 0 when we observe class 0.
- It will have a maximum value of 1 where all of the training points are predicted perfectly.

Bayesian model for binary responses/targets

- Here, our objective is to choose a function of $\mathbf{x}^{(i)}$ and $\mathbf{w}$, denoted by $f(\mathbf{x}^{(i)}; \mathbf{w})$ that produces a probability.
- We may take a linear function $f(\mathbf{x}^{(i)}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$ and then pass the result through a second function that would squashes its output to ensure it produces a valid probability –> Sigmoid function.

$$P(Y^{(i)} = 1|\mathbf{x}^{(i)}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}^{(i)})}$$

- The above expression ensures us the probability that $Y^{(i)} = 1$.



Bayesian model for binary responses/targets

- Since $Y^{(i)} \in \{1, 0\}$, $P(Y^{(i)} = 0|\mathbf{x}^{(i)}, \mathbf{w})$ can be written as

$$P(Y^{(i)} = 0|\mathbf{x}^{(i)}, \mathbf{w}) = 1 - P(Y^{(i)} = 1|\mathbf{x}^{(i)}, \mathbf{w})$$

- By using $P(Y^{(i)} = 1|\mathbf{x}^{(i)}, \mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}$, we have

$$\begin{aligned} P(Y^{(i)} &= 0|\mathbf{x}^{(i)}, \mathbf{w}) = 1 - \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}^{(i)})} \\ &= \frac{\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}{1 + \exp(-\mathbf{w}^T\mathbf{x}^{(i)})} \end{aligned}$$

Bayesian model for binary responses/targets

- A single expression can be written as

$$P(Y^{(i)} = y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) = \Big[P(Y^{(i)} = 1|\mathbf{x}^{(i)}, \mathbf{w})\Big]^{y^{(i)}} \Big[P(Y^{(i)} = 0|\mathbf{x}^{(i)}, \mathbf{w})\Big]^{1-y^{(i)}}$$

- Next, the likelihood function can also be rewritten as

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \prod_{i=1}^{N} P(Y^{(i)} = y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) \\ &= \prod_{i=1}^{N} \Big[P(Y^{(i)} = 1|\mathbf{x}^{(i)}, \mathbf{w})\Big]^{y^{(i)}} \Big[P(Y^{(i)} = 0|\mathbf{x}^{(i)}, \mathbf{w})\Big]^{1-y^{(i)}} \\ &= \prod_{i=1}^{N} \left[\frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}^{(i)})}\right]^{y^{(i)}} \left[\frac{\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}{1 + \exp(-\mathbf{w}^T\mathbf{x}^{(i)})}\right]^{1-y^{(i)}} \end{aligned}$$

Bayesian model for binary responses/targets

- The posteriori distribution is given by

$$f(\mathbf{w}|\mathbf{y},\mathbf{X}) = \frac{f(\mathbf{y}|\mathbf{X},\mathbf{w})f(\mathbf{w})}{f(\mathbf{y}|\mathbf{X})},$$

  where $f(\mathbf{y}|\mathbf{X}) = \int f(\mathbf{y}|\mathbf{X},\mathbf{w})f(\mathbf{w})d\mathbf{w}$, and the prior distribution of $\mathbf{w}$ can be assumed to Gaussian distributed $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

- Moreover, in the above, $f(\mathbf{y}|\mathbf{X},\mathbf{w})$ is the likelihood function that we derived previously as

$$f(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_{i=1}^{N} \left[\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}\right]^{y^{(i)}} \left[\frac{\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}{1+\exp(-\mathbf{w}^T\mathbf{x}^{(i)})}\right]^{1-y^{(i)}}$$

- With the posterior density, the response of new objects can be predicted by taking an expectation with respect to posterior density:

$$P(y_{\text{new}} = 1|x_{\text{new}},\mathbf{X},\mathbf{y}) = \mathbb{E}_{f(\mathbf{w}|\mathbf{X},\mathbf{y})}\left[\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}_{\text{new}})}\right],$$

where $f(\mathbf{w}|\mathbf{X},\mathbf{y})$ is the conditional posterior density of $\mathbf{w}$.