Task – 2

Sub Task – 1

- Prakhar Tripathi

A detailed mathematical explanation of the Gaussian Discriminative Analysis model describing the principles, assumptions, and equations involved and explaining how the model learns from the data and makes predictions.

Contents

Generative Learning Algorithms

In general models train on p(y| x; θ), the conditional distribution of y over x. For instance, logistic regression modelled p(y| x; θ), as $\mathrm{h}_\theta(\mathrm{x}) = \mathrm{g}(\theta^{\mathrm{T}}\mathrm{x})$ where g is the sigmoid function.

Lets take an example where we have to distinguish between car(y=0) and aeroplane(y=1) based on some features of vehicles. If we are given a training set, an algorithm like Logistic Regression tries to find a straight line that is a decision boundary that separates the regions of cars and aeroplanes on the graph.

But let's see a different approach where we can build a model of what cars look like. Then looking at aeroplane, we can build a separate model of what aeroplanes look like. Finally, to classify a new vehicle, we can match the new vehicle against the cars model, and match it against the elephant model, to see whether the new vehicle looks more like the aeroplanes or more like the cars we had seen in the training set.

Algorithms that try to learn p(y| x) directly (such as logistic regression), or algorithms are called discriminative learning algorithms. Here, well talk about algorithms that instead try to model p(x| y) and p(y). These algorithms are called generative learning algorithms. For instance, if y indicates whether an example is a car (0) or an aeroplane (1), then p(x|y = 0) models the distribution of car features, and p(x|y = 1) models the distribution of aeroplanes features.

After modelling p(y) (called the class priors) and p(x| y), our algorithm can then use Bayes rule to derive the posterior distribution on y given x:

$$\mathrm{p(y|x)} = \frac{\mathrm{p(x|y)p(y)}}{\mathrm{p(x)}}$$

Here, the denominator is given by p(x) = p(x| y = 1)p(y = 1) + p(x| y = 0)p(y = 0)

Actually, if were calculating p(y| x) in order to make a prediction, then we don't actually need to calculate the denominator, since

$$\begin{aligned} arg\underbrace{\mathrm{max}}_{\mathrm{y}} p(y|x) &= arg\underbrace{\mathrm{max}}_{\mathrm{y}} \frac{p(x|y)p(y)}{p(x)} \\ &= \mathrm{arg}\ \underbrace{\mathrm{max}}_{\mathrm{y}} \mathrm{p(x|y)p(y)} \end{aligned}$$

Gaussian Discriminative Analysis

In this model, well assume that p(x| y) is distributed according to a multivariate normal distribution. Let’s talk briefly about the properties of multivariate normal distributions before moving on to the GDA model itself.

The multivariate Normal Distribution

The multivariate normal distribution in d-dimensions, also called the multi variate Gaussian distribution, is parameterized by a mean vector $\mu \in \mathbb{R}^{\mathbf{d}}$ and a covariance matrix $\Sigma \in \mathbb{R}^{\mathbf{dxd}}$ where $\Sigma \geq 0$ is symmetric and positive semi-definite. Also written $\aleph(\mu, \Sigma)$, its density is given by

$$\mathrm{p}(\mathrm{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{\mathrm{d}/2}|\Sigma|} \exp\left(-\frac{1}{2}(\mathrm{x} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathrm{x} - \mu)\right)$$

In the above equation, $|\Sigma|$ denotes the determinant of Σ.

The Gaussian Discriminant Analysis Model

When we have a classification problem in which the input features x are continuous-valued random variables, we can then use the Gaussian Discriminant Analysis (GDA) model, which models p(x| y) using a multivariate normal distribution. The model is:

$$y \sim Bernoulli(\phi)$$
$$\mathrm{x}|\mathrm{y} = 0 \sim \aleph(\mu_0, \Sigma)$$
$$\mathrm{x}|\mathrm{y} = 1 \sim \aleph(\mu_1, \Sigma)$$

Writing out the distributions this is

$$\mathrm{p}(\mathrm{y}) = \phi^{\mathrm{y}}(1 - \phi)^{1-\mathrm{y}}$$

$$\mathrm{p}(\mathrm{x}|\mathrm{y} = 0) = \frac{1}{(2\pi)^{\mathrm{d}/2}|\Sigma|} \exp\left(-\frac{1}{2}(\mathrm{x} - \mu_0)^{\mathrm{T}}\Sigma^{-1}(\mathrm{x} - \mu_0)\right)$$

$$\mathrm{p}(\mathrm{x}|\mathrm{y} = 1) = \frac{1}{(2\pi)^{\mathrm{d}/2}|\Sigma|} \exp\left(-\frac{1}{2}(\mathrm{x} - \mu_1)^{\mathrm{T}}\Sigma^{-1}(\mathrm{x} - \mu_1)\right)$$

The log-likelihood of the data is given by

$$\mathrm{l}(\phi,\mu_0,\mu_1,\Sigma\ ) = \log\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{x}^{(\mathrm{i})},\mathrm{y}^{(\mathrm{i})};\phi,\mu_0,\mu_1,\Sigma)$$

$$= \log\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{x}^{(\mathrm{i})}|\mathrm{y}^{(\mathrm{i})};\phi,\mu_0,\mu_1,\Sigma)\ \mathrm{p}(\mathrm{y}^{(\mathrm{i})};\phi)$$

By maximizing with respect to the parameters, we find the maximum like likelihood estimate of the parameters:

$$\phi = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} 1\{\mathrm{y}^{(i)} = 1\}$$

$$\mu_0 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} 1(\{\mathrm{y}^{(\mathrm{i})} = 0\}\mathrm{x}^{(\mathrm{i})}}{1\{\mathrm{y}^{(\mathrm{i})} = 0\}}$$

$$\mu_1 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} 1(\{\mathrm{y}^{(\mathrm{i})} = 1\}\mathrm{x}^{(\mathrm{i})}}{1\{\mathrm{y}^{(\mathrm{i})} = 1\}}$$

$$\Sigma = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x}^{(\mathrm{i})}\text{-}\mu_{\mathrm{y}^{(\mathrm{i})}})(\mathrm{x}^{(\mathrm{i})}\text{-}\mu_{\mathrm{y}^{(\mathrm{i})}})^{\mathrm{T}}$$

Prediction and Output

The two models can be built by these values of mu and variances. Then each data’s probability is calculated in both classes. Then the one with the bigger value is allotted the class.