Lecture 7 - Generative Learning

So far, we have covered discriminative algorithms. This is because they directly model P(y—X), only caring about the chance of the data being correct given the X provided. However, now we want to focus on X. We want to model P(X, y) i.e what is the chance of getting X and y.

P(X, y) = $P(X|y) * P(y)$ is what we want to model.

When we want to classify a new example, we use $P(y|X) = \frac{P(X|y)*P(y)}{P(X)}$ (Bayes theorem)
This expression is equal to $P(y|X) = \frac{P(X|y)*P(y)}{P(X|y=0)*P(y=0)+P(X|y=1)*P(y=1)}$, which is equivalent because the denominator is all the possible ways to get X (regardless of which class it belongs to).

If we want to predict a class (for binary classification): $\hat{y}$ = arg max $P(y|X)$ for each different class e.g compare the probability of getting a class given an input for all classes, and then select the class with the highest probability.

Note: $P(y)$ is the proportion of y's which are 1's from the dataset, and $P(X)$ is the chance of encountering a certain input example in the dataset.

Gaussian Discriminative Analysis (GDA)

Take several modelling assumptions for GDA:

$y \sim Bern(\phi)$
$X \mid y \sim \mathcal{N}(\mu_0, \Sigma)$ (There is a continuous distribution of $X_1, X_2....X_d$ which follows bivariate normal)
$X \mid y \sim \mathcal{N}(\mu_1, \Sigma)$ (Each class has the same shape (Covariance) but different mean)

Based on these assumptions, it follows that:

$P(y) = \phi^y(1-\phi)^{1-y}$
$P(X \mid y = 0) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^2} \exp(-\frac{1}{2}(X-\mu_0)^T\Sigma^{-1}(X-\mu_0))$
$P(X \mid y = 0) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^2} \exp(-\frac{1}{2}(X-\mu_1)^T\Sigma^{-1}(X-\mu_1))$

Model parameters

$\phi, \mu_0, \mu_1, \Sigma$ - if we can find parameters that fit our data modeling process as well as possible, we can find how X and y are distributed.

Log Likelihood

$l(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} P(X^{(i)}, y^{(y)})$

The likelihood is multiplying all the chances of finding each X, y example given our current guess at the parameters. We multiply based on the IID assumption (Independent identically distributed) - as we multiply probabilities if they both occurred (think of a probability tree). The reason we take the log is because multiplying a bunch of very small probabilities together will quickly result in underflow. However log coverts products to sums, and so this does not happen. Additionally, as log(x) is a monotonic (a < b ∴ log(a) < log(b)) transformation, maximizing the log-likelihood is equivalent to maximizing the original likelihood, meaning that the parameter values that maximize one will also maximize the other.
By maximizing the likelihood function (differentiation with respect to each parameter) for each parameter, we maximize the chance of each X, y (taken from our dataset) appearing - i.e we are as closely as possible fitting the distribution that our data was generated by (based on our previous assumptions)

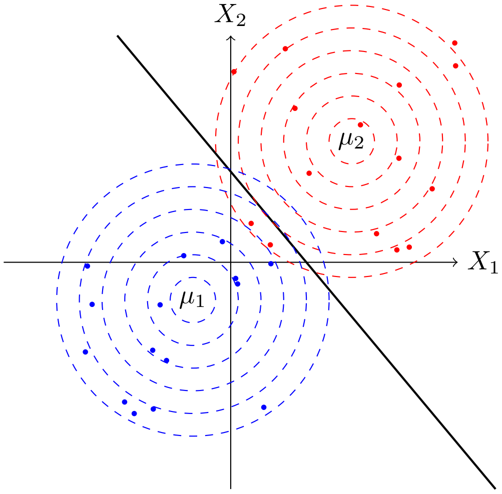
For this example, there is a single closed form solution, achievable by setting the derivative = 0. It can be proved that $P(y = 1 \mid X) = \frac{1}{1+e^{-\theta^T X}}$ Where θ depends on only $\phi, \mu_0, \mu_1, \Sigma$. This is the logistic function.
NOTE: For the following examples, 1{condition} notation means 1 if condition true, 0 if condition false.

$$\hat{\phi} = \frac{1}{n} \sum 1\{y^{(i)} = 1\} \text{ - simply the proportion of samples which are a 1 (num 1's / n)}$$

$$\hat{\mu_0} = \frac{\sum 1\{y^{(i)} = 0\} x^{(i)}}{\sum 1\{y^{(i)} = 0\}} \text{ - The average of all the X's which produced y = 1}$$

$$\hat{\mu_0} = \frac{\sum 1\{y^{(i)} = 1\} x^{(i)}}{\sum 1\{y^{(i)} = 1\}} \text{ - The average of all the X's which produced y = 1}$$

$$\hat{\Sigma} = \frac{1}{n} \sum (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$ - How far each X example lies from its own mean (μ_0, μ_1 must be calculated before this).



This graph represents the data generation of 2 classes - red and blue, for a length 2 X vector. The diagonal line is the decision boundary, where $P(X \mid y = 1) = P(X \mid y = 0)$, i.e the points where the chance of a certain X appearing is the same regardless of the class.
We are trying to find the $\mu_0, \mu 1, \phi, \Sigma$ (the dotted distributions), which best fit our data points (the dots). Bear in mind that this would be a d directional space for a X vector length d.