Stanford CS 229 - Machine Learning - Lecture 4

Giovani da Silva

June 2021

1 Generative learning algorithm

Instead of trying to draw a line to separate two types, we can try to learn them separately and after that build one model for each one, then we can use that to build a model to our training set.

Discriminative learning algorithms try to learn p(y—x) directly (such as logistic regression), or algorithms that try to learn mappings directly from the space of inputs X to the labels $\{0, 1\}$, (such as the perceptron algorithm) are called generative learning algorithms instead try to model p(x—y) (and p(y)). The interpretation of this is that a generative model builds a probabilistic model for what the features looks like, conditioned on the class label.

After modelling p(y) and p(x—y) we can use nayve bayes rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$
$where : p(x) = p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)$

1.1 Gaussian Discriminant Analysis model

When we classification problem in which the input features x are continuous-valued random variables.which models p(x—y) using a multivariate normal distribution.

Writing out the distributions, this is:

$$\begin{aligned} p(y) &= \phi^y(1-\phi)^{1-y} \\ p(x|y=0) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\ p(x|y=1) &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) \end{aligned}$$

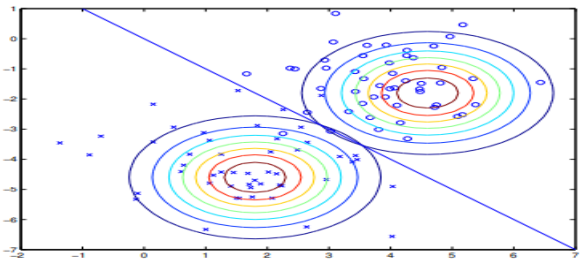
Here, the parameters of our model are ϕ, Σ, μ_0 and μ_1. (Note that while there're two different mean vectors μ_0 and μ_1, this model is usually applied using only one covariance matrix Σ.) The log-likelihood of the data is given by

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi). \end{aligned}$$

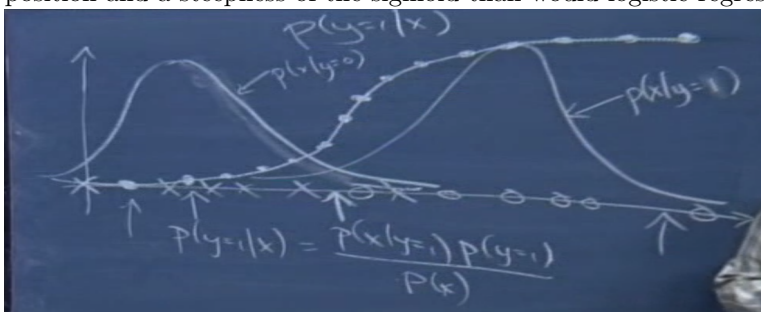
By maximizing ℓ with respect to the parameters, we find the maximum likelihood estimate of the parameters (see problem set 1) to be:

$$\begin{aligned} \phi &= \frac{1}{n}\sum_{i=1}^{n} 1\{y^{(i)}=1\} \\ \mu_0 &= \frac{\sum_{i=1}^{n} 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=0\}} \\ \mu_1 &= \frac{\sum_{i=1}^{n} 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=1\}} \\ \Sigma &= \frac{1}{n}\sum_{i=1}^{n}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T. \end{aligned}$$

Pictorially, what the algorithm is doing can be seen in as follows:



GDA has a relation with logistic regression:
When you make the assumptions under the Gaussian discriminant analysis model, that P(x—y) is Gaussian, when you go back and compute whatP(y—x) is, you actually get back exactly the same sigmoid function. t turns out the key difference is that Gaussian discriminant analysis will end up choosing a different position and a steepness of the sigmoid than would logistic regression.



Gaussian discriminant analysis makes a much stronger assumption that X given Y is Gaussian, and so when this assumption is true, , if you plot the data, and if X given Y is, indeed, approximately Gaussian, then if you make this assumption, explicit to the algorithm, then the algorithm will do better because it's as if the algorithm is making use of more information about the data. The algorithm knows that the data is Gaussian. And so if the Gaussian assumption holds or roughly holds, then Gaussian discriminant analysis may do better than logistic regression.

If you're actually not sure what X given Y is, then logistic regression is going to be better. Use logistic regression, and if turns out the data was actually Poisson,for example, then logistic regression will still do perfectly fine but if you assumed it was Gaussian, then the algorithm may go off.

It turns out the real advantage of generative learning algorithms is often that it requires less data, and, in particular, data is never really exactly Gaus-

sian. Because data is often approximately Gaussian; it's never exactly Gaussian. And it turns out, generative learning algorithms often do surprisingly well even when these modeling assumptions are not met, but one other tradeoff is that by making stronger assumptions about the data, Gaussian discriminant analysis often needs less data in order to fit, like, an okay model, even if there's less training data;
Whereas, in contrast, logistic regression by making less assumption is more robust to your modeling assumptions because, you're making less assumptions, but sometimes it takes a slightly larger training set to fit than Gaussian discriminant analysis.

1.2 Naive Bayes

The x's are discrete-valueted In a Naive Bayes algorithm, we're going to make a very strong assumption on p(x—y) . We are going to assume that the $X_i's$ are conditionally independent given Y.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

1.2.1 Laplace smoothing

Laplace smoothing is a smoothing technique that handles the problem of zero probability in Naïve Bayes.

$$P(w|positive) = \frac{numberwithw'andy=positive+\alpha}{N+\alpha*K}$$

or

$$\begin{aligned}
\phi_{j|y=1} &= \frac{1+\sum_{i=1}^{n} 1\{x_j^{(i)}=1 \wedge y^{(i)}=1\}}{2+\sum_{i=1}^{n} 1\{y^{(i)}=1\}} \\
\phi_{j|y=0} &= \frac{1+\sum_{i=1}^{n} 1\{x_j^{(i)}=1 \wedge y^{(i)}=0\}}{2+\sum_{i=1}^{n} 1\{y^{(i)}=0\}}
\end{aligned}$$