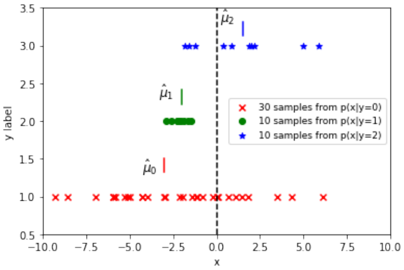
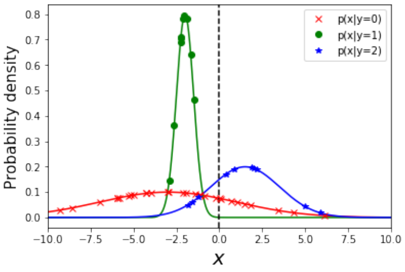
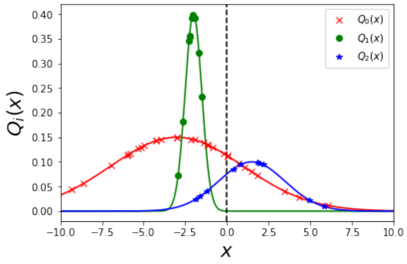
CS 189 Introduction to Machine Learning
Spring 2019 Jonathan Shewchuk
DIS4

1 Gaussian Discriminant Analysis

We have N iid samples $\{(X_n, Y_n)\}_{n=1}^N$ with values $\{(x_n, y_n)\}_{n=1}^N$, where $x_n \in \mathbb{R}$ is an observable and $y_n \in \{0,1,2\}$ is the class to which the sample belongs. We'll denote by N_i the number of samples that belong to class $Y = i$. We have plotted the samples in the figure to the left. You want to build a classifier such that you can predict the class of new unlabeled samples $X = x$. You have been told that the conditional probabilities $p(x|y)$ are Gaussian distributions.







(a) How would you use Maximum Likelihood Estimation (MLE) to estimate the probabilities $p(X|Y)$ and $\pi_i = p(Y = i)$ from the samples?

Solution: Using the given information, we use gaussian distributions as our models $p_{\mu_i,\sigma_i}(X = x|Y = i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$. By looking at the samples, the in-class variances seem substantially different, so we estimate a individual σ_i per class. The MLE in this case is the usual sample mean and sample variance (no need to re-derive it, but make sure to understand why those are the MLE)

$$\hat{\mu}_i = \frac{1}{N_i} \sum_{n:y_i=i} x_n; \qquad \hat{\sigma}_i^2 = \frac{1}{N_i} \sum_{n:y_n=i} (x_n - \hat{\mu}_i)^2$$

The prior distributions are the proportion of samples from each class: $\pi_i = p(Y = i) = \frac{N_i}{\sum_{j=0}^2 N_j}$.

(b) How would you use these probabilities to derive the Bayes decision rule? What equations are satisfied by the points in the decision boundary $r^*(x)$? Leave the solution in terms of $Q_0(x), Q_1(x), Q_2(x)$, where

$$e^{Q_i(x)} = \sqrt{(2\pi)} p(X = x|Y = i) p(Y = i) = \frac{\pi_i}{\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$

.

Solution: The optimal Bayes rule assigns to any sample $X = x$ the class that has largest probability $\arg\max_i p(Y = i|X = x)$. We can use a plot like the one on the right to visually see which class is more likely for any value of the sample X, under our model. More concretely:

$$\arg\max_i p(Y = i|X = x) = \arg\max_i \frac{p(X = x|Y = i)p(Y = i)}{\sum_j p(X = x|Y = j)p(Y = j)} = \arg\max_i \frac{e^{Q_i(x)}}{\sum_j e^{Q_j(x)}}$$

The function being optimized is called the soft-max function.

On each side of the decision boundary the most likely class is different. Let's say that on one side $Y = i$ is the most likely and on the other one $Y = j$ is the most likely. In the first side $e^{Q_i(x)} \geq e^{Q_k(x)}\ \forall k$, and on the other side $e^{Q_j(x)} \geq e^{Q_k(x)}\ \forall k$, therefore at the boundary both sets of inequalities are satisfied. In particular, we can see it implies $Q_i(x) = Q_j(x)$! Therefore all points in the decision boundary will satisfy this for some values of $i,\ j : i \neq j$. The same conclusion can be reached by looking at the right graph.

(c) What do you observe about the region of values of X where the label $Y = 0$ is assigned? Could you express this region with a set of inequalities?

Solution: We see that the decision region corresponding to $Y = 0$ consists of three disjoint non-empty components, approximately $[\infty, -2.8]$, $[-1.6, 0.9]$, $[5.4, \infty]$. It is thus disconnected, and hence non-convex. Since it is non-convex it cannot be expressed as the intersection of a set of inequalities.

(d) You receive a new unlabeled sample $X = 0$, what class would you assign to it? Is it the class which mean is closest?

Solution: It will go to the class $Y = 0$, but in fact both means μ_1 and μ_2 are closer to 0!!!

(e) What would have happen if you used Linear Discriminant Analysis and assumed uniform priors?

Solution: If you used LDA you would assume the conditional probabilities of each class have the same variance. In this case, under uniform priors, the assigned class is always the one that has the closest mean to the new sample X.

(f) *Bonus question: Is it possible that there is a certain class $y = i$ for which there is no x such that the Bayes decision rule picks this class i*

Solution: *Bonus answer: If arbitrary priors are allowed, you can easily see that you can always make one class less likely everywhere*

2 Maximum Likelihood Estimation for reliability testing

Suppose we are reliability testing n units taken randomly from a population of identical appliances. We want to estimate the mean failure time of the population. We assume the failure times come from an exponential distribution with parameter $\lambda > 0$, whose probability density function is $f(t) = \lambda e^{-\lambda t}$ (on the domain $t \geq 0$).

(a) In an ideal (but impractical) scenario, we run the units until they all fail. The failure time $T_1, T_2, \ldots, T_n$ for units $1, 2, \ldots, n$ are observed to be $t_1, t_2, \ldots, t_n$.

Formulate the likelihood function $\mathcal{L}(\lambda; t_1, \ldots, t_n)$ for our data. Then find the maximum likelihood estimate $\hat{\lambda}$ for the distribution's parameter. (Remember that it's equivalent, and usually easier, to optimize the log-likelihood)

Solution:

$$\begin{aligned}
\mathcal{L}(\lambda; t_1, \ldots, t_n) &= \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i} \\
\ln \mathcal{L}(\lambda) &= n \ln \lambda - \lambda \sum_{i=1}^{n} t_i \\
\frac{\partial}{\partial \lambda} \ln \mathcal{L}(\lambda) &= \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0 \\
\hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} t_i}
\end{aligned}$$

(b) In a more realistic scenario, we run the units for a fixed time h. The failure time for $T_1, T_2, \ldots, T_r$ are observed to be $t_1, t_2, \ldots, t_r$, where $0 \leq r \leq n$. The remaining $n-r$ units survive the entire time h without failing. Let's find the maximum likelihood estimate $\hat{\lambda}$ for our model distribution parameters! To do so:

(a) What is the probability that a unit will not fail during time h?

(b) Write the new likelihood function $\mathcal{L}(\lambda; h, n, r, t_1, \ldots, t_r)$.

(c) Optimize to find the MLE estimate, and give it a physical interpretation.

Solution:

(a) $P(T > h) = 1 - P(T \leq h) = 1 - \int_{t=0}^{h} f(t)dt = 1 - \left[e^{-\lambda t}\right]_0^h = 1 - (1 - e^{-\lambda h}) = e^{-\lambda h}$

(b)

$$\begin{aligned}
\mathcal{L}(\lambda; n, h, r, t_1, \ldots, t_r) &= P(T_1 = t_1, \ldots, T_r = t_r, T_{i>r} > h; \lambda) \\
&= \left(\prod_{i=1}^{r} f(t_i)\right) P(t > h)^{n-r} \\
&= \left(\prod_{i=1}^{r} \lambda e^{-\lambda t_i}\right) \left(e^{-\lambda h}\right)^{n-r} \\
&= \lambda^r e^{-\lambda \sum_{i=1}^{r} t_i} e^{-\lambda (n-r) h}
\end{aligned}$$

Note: The wording of the question specifies which units failed, therefore we shouldn't add the "n choose r" type of coefficient. Still, check in next question that even if you weren't told which were the units that failed when, the MLE solution would be the same!

(c)

$$
\begin{aligned}
\ln \mathcal{L}(\lambda) &= r \ln \lambda - \lambda \sum_{i=1}^{r} t_i - \lambda(n-r)h \\
\frac{\partial}{\partial \lambda} \ln \mathcal{L}(\lambda) &= \frac{r}{\lambda} - \sum_{i=1}^{r} t_i - (n-r)h = 0 \\
\hat{\lambda} &= \frac{r}{\sum_{i=1}^{r} t_i + (n-r)h}
\end{aligned}
$$

We can interpret $\hat{\lambda}$ will be the number of observed failures divided by the sum of unit test times.