Mid-term review
ECE 271A
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos Fall 2008

1. Consider a classification problem with two Gaussian classes

$$P_{\mathbf{X}|Y}(\mathbf{x}|i) = \mathcal{G}(\mathbf{x}, \mu_i, \Sigma), \quad i \in \{0, 1\}$$

of equal probability

$$P_Y(i) = 1/2.$$

In class, we have considered the BDR solution to this problem. This consists of estimating the parameters of the Gaussian classes and then plugging on the BDR to obtain the decision boundary. Here we will consider an alternative solution, that works directly on the class posteriors.

a) Show that the posterior probability for class 1 is of the form (the posterior for class 0 is $1 - P_{Y|\mathbf{x}}(1|\mathbf{x})$)

$$P_{Y|\mathbf{X}}(1|\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{t}}} \tag{1}$$

where $\mathbf{t}^T = [\mathbf{x}^T 1]$. What is the vector $\mathbf{w}$?

b) Show that an iid sample $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ has posterior probability

$$P_{\mathbf{Y}|\mathbf{X}}(\mathcal{D}_y|\mathcal{D}_x) = \prod_{i=1}^{n} P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) \tag{2}$$

with

$$P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) = \left(\frac{1}{1 + e^{-\mathbf{w}^T \mathbf{t}_i}}\right)^{y_i} \left(\frac{e^{-\mathbf{w}^T \mathbf{t}_i}}{1 + e^{-\mathbf{w}^T \mathbf{t}_i}}\right)^{1-y_i}, \tag{3}$$

where $\mathcal{D}_y = \{y_1, \ldots, y_n\}$ and $\mathcal{D}_x = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

c) We can now learn the classification boundary, by learning the posterior probabilities with standard maximum likelihood estimation. For example we can solve for $\mathbf{w}^\star$ such that

$$\mathbf{w}^\star = \arg\max_{\mathbf{w}} P_{\mathbf{Y}|\mathbf{X}}(\mathcal{D}_y|\mathcal{D}_x).$$

Show that $\mathbf{w}^\star$ must satisfy the condition

$$\sum_i y_i \mathbf{t}_i = \sum_i \frac{1}{1 + e^{-\mathbf{w}^{\star T} \mathbf{t}_i}} \mathbf{t}_i$$

d) Can you guess what the optimal $\mathbf{w}^\star$ is? When n goes to infinity the condition above converges to (dividing by n on each side)

$$E_{Y,\mathbf{X}}[Y\mathbf{T}] = E_{\mathbf{X}}\left[\frac{1}{1 + e^{-\mathbf{w}^{\star T} \mathbf{T}}} \mathbf{T}\right]$$

with $\mathbf{T}^T = [\mathbf{X}^T 1]$. Show that this condition holds for your guess.

2. Consider a two dimensional classification problem with two Gaussian classes

$$P_{\mathbf{X}|Y}(\mathbf{x}|i) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_i)}, \quad i \in \{0,1\}$$

of identical covariance $\mathbf{\Sigma} = \sigma^2\mathbf{I}$. For all problems assume the "0-1" loss function.

a) If the classes have means

$$\mu_0 = -\mu_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

and equal prior probabilities, $P_Y(0) = P_Y(1)$, what is the Bayes decision rule for this problem?

b) What are the marginal distributions for the features x_1 and x_2 for each class? In particular

1. derive expressions for the class-conditional densities $P_{X_1|Y}(x_1|i)$ and $P_{X_2|Y}(x_2|i)$ for $i \in \{0,1\}$, where $\mathbf{x} = (x_1, x_2)^T$.
2. plot a sketch of the two densities associated with class $Y = 0$ and a sketch of the two densities associated with class $Y = 1$.
3. determine which feature is most discriminant.

c) A linear transformation of the form

$$\mathbf{z} = \mathbf{\Gamma}\mathbf{x}$$

was applied to the data, where $\mathbf{\Gamma}$ is a 2×2 matrix. The decision boundary associated with the BDR is now the hyperplane of normal $\mathbf{w} = (1/\sqrt{2}, -1/\sqrt{2})^T$ which passes through the origin.

1. determine the matrix $\mathbf{\Gamma}$
2. What would happen if the the prior probability of class 0 was increased after the transformation? Here it suffices to give a qualitative answer, i.e. simply say what would happen to the hyperplane.

d) Classification assumes that we know the label for each training point. Sometimes we do not, i.e. we have the $\mathbf{x}_i$ but no y_i. The classification ideas can, however, be extended to this problem, resulting in what is usually called a *clustering* algorithm. The main idea is the following:

- assume two Gaussian classes of identity covariance $\mathbf{\Sigma}_i = \mathbf{I}$;
- start from a random estimate for the means μ_i, e.g. pick two points from the training set at random, and a $P_Y(0) = P_Y(1)$ estimate for the class probabilities
- perform K iterations of the following steps:
    1. classify the points using the BDR. That is, assign each point $\mathbf{x}_i$ to class $Y = 0$ or $Y = 1$ using the current Gaussian parameter estimates and the BDR;
    2. update the Gaussian parameters using the new point assignments. That is use all the points assigned to class $Y = 0$ to recompute the mean and probability of class $Y = 0$, and all assigned to class $Y = 1$ to recompute the parameters of this class.

Denoting the class parameters estimated at iteration k by $P_Y^k(i)$ and μ_i^k, and the points assigned to class $Y = i$ in this iteration by $D_i^k = \{\mathbf{x}_{i,1}^k, \ldots, \mathbf{x}_{i,n}^k\}$, answer the following questions.

1. what are the equations for point assignment? That is, given the parameters $P_Y^k(i)$, and μ_i^k, and a point $\mathbf{x}$, how would you determine the point's class?
2. what are the parameter update equations? That is, given the set $D_i^k = \{\mathbf{x}_{i,1}^k, \ldots, \mathbf{x}_{i,n}^k\}$ of points assigned to class $Y = i$, what are the class parameters for the next iteration, i.e.

$$P_Y^{k+1}(i) \text{ and } \mu_i^{k+1}?$$

3. consider the following 4-point training set

$$\{(-1, 0.9)^T, (1, 0.8)^T, (1.1, 1)^T, (-1.2, 1.1)^T\}$$

and let the initial parameter estimates be $\mu_0 = (-1.2, 1.1)^T$ and $\mu_1 = (-1, 0.9)^T$. Run the algorithm for 3 iterations. For each iteration plot a sketch with the points and the BDR. Comment on whether the algorithm was successful in separating the classes at the end of each iteration.

Note: you do not necessarily have to do all the precise computations. A "graphical" solution is sufficient. If you find some ambiguity that would require precise computation to disambiguate, e.g. you need to find out if $a < b$ and a is very similar to b, feel free to just assume one way or the other, and write that down. We will not take points for that, as along as it is within reason, of course.