10-701/15-781, Machine Learning: Homework 3 Solution

1 Learning Theory [25pt, Ni Lao]

1.1 Gaussian Bayes Model [15 pt]

ID	**a**	**b**	**c**	**d**	**e**	**f**
No. features	1	1	2	2	2	2
Shared Covariance Matrix?	Y	N	Y	Y	N	N
Naive Bayes?	-	-	Y	N	Y	N
No. parameters	1*2+1 +1=4	1*2+1*2 +1=5	2*2+2 +1=7	2*2+3 +1=8	2*2+2*2 +1=9	2*2+3*2 +1=11
VC dimension	2	3	3	3	>=5	>=6

The decision boundary of a two class Gaussian model is

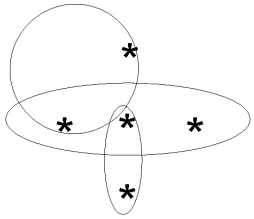
$$\log|\pi_1| - \log|\Sigma_1| - (\mathrm{x}-\mu_1)'\Sigma_1^{-1}(\mathrm{x}-\mu_1) = \log|\pi_2| - \log|\Sigma_2| - (\mathrm{x}-\mu_2)'\Sigma_2^{-1}(\mathrm{x}-\mu_2) \quad (1)$$

It can be rearranged as

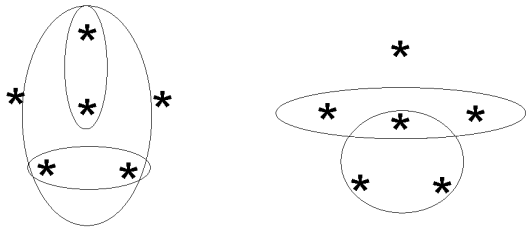
$$\alpha + \beta'\mathrm{x} + \mathrm{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\mathrm{x} = 0 \quad (2)$$

When the covariance matrix is shared ($\Sigma_1 = \Sigma_2$), we get a linear boundary, which can shatter 2 points in 1 dimensional space and 3 points in 2 dimensional space.

When the covariance matrix is not shared, the boundary is an ellipsoid. If the model is naive Bayes then the axis of ellipsoid must be parallel to the basis of the space. As demonstrated below, we can separate out any 2 points out of 5. Separating out any single point is trivial. If the model



has no naive Bayes assumption, we can shatter 6 points. As demonstrated on the left below, we can separate out any 2 points, and on the right below, we can separate out any 3 points.



Alternative solution (provided by Chi Song)

For case f, Eq.(2) has 6 parameters, but actually only 5 of them are free (there is a family of parameter settings that produces the same curve). Therefore, we can always fit the curve to any arbitrary 5 points.

Suppose we have 6 data points $x_1, \ldots x_6$. We can always fit the curve to $x'_1, \ldots x'_5$, the slight perturbation of $x_1, \ldots x_5$, so that $x_1, \ldots x_5$ are correctly classified. If x_6 is already correctly classified, then we are done. If not, we can reverse the sign of the quadratic form and do a further perturbation to $x_1, \ldots x_5$, thus force them to be classified correctly. Therefore $VC \geq 6$.

For case e, the quadratic form has one less number of freedom. Therefore $VC \geq 5$.

1.2 VC dimension and Effective Number of Hypothesis [3 pt]

When $m \leq VC(H)$, $\max_{\mathcal{D}} N_{eff}(H, \mathcal{D}) = 2^m$.
When $m > VC(H)$, $\max_{\mathcal{D}} N_{eff}(H, \mathcal{D}) < 2^m$.

1.3 [2 pt]

$\log_2 N_{eff}(H, \mathcal{D})$

1.4 Linear Regression Model [3 pt]

$$R(\lambda)/n = \frac{p}{n}\sigma^2 + \sigma^2, \tag{3}$$

1.5 [2 pt]

We need p samples, which is the VC dimension of a linear model minus 1.