Machine Learning Problemset 1

Aimee Barciauskas

29 January 2016

Problemset 1

1. Consider the binary classification problem with a priori probabilities $P\{Y = 1\} = P\{Y = 0\} = \frac{1}{2}$ and class-conditional densities $f_0(x) = f(x|Y = 0)$ and $f_1(x) = f(x|Y = 1)$ on $X = \mathbb{R}_d$. Prove that the Bayes risk equals:

$$R^* = \frac{1}{2} - \frac{1}{4}\int \big|f_1(x) - f_0(x)\big|dx$$

We know that:

$$R^* = \int min(\eta(x), 1 - \eta(x))dx$$

$$\eta(x) = (1 - \frac{1}{2})f_1(x)$$

$$1 - \eta(x) = \frac{1}{2}f_0(x)$$

Substituting for $\eta(x)$ gives:

$$R^* = \int min((1 - \frac{1}{2})f_1(x), \frac{1}{2}f_0(x))dx$$

$$= \int min(f_1(x) - \frac{1}{2}f_1(x), \frac{1}{2}f_0(x))dx$$

$$= \int min(\frac{1}{2}f_1(x), \frac{1}{2}f_0(x))dx$$

$$= \frac{1}{2}\int min(f_1(x), f_0(x))dx$$

The minimum of two functions can be expressed:

$$min(f_1(x), f_0(x)) = \frac{1}{2}[f_1(x) + f_0(x) - \big|f_1(x) - f_0(x)\big|]$$

Substituting this equality:

$$R^* = \frac{1}{2}\int \frac{1}{2}[f_1(x) + f_0(x) - \big|f_1(x) - f_0(x)\big|]dx$$

$$= \frac{1}{4}\Big[\int f_1(x)dx + \int f_0(x)dx - \int \big|f_1(x) - f_0(x)\big|dx\Big]$$

Since $f_0(x)$ and $f_1(x)$ are probability density functions, the first two terms integrate to 1, so the above can be reduced to:

$$= \frac{1}{2} - \frac{1}{4}\int \big|f_1(x) - f_0(x)\big|dx$$

2. Consider a binary classification problem in which both class-conditional densities are multivariate normal of the form

$$f_i(x) = \frac{1}{\sqrt{2\pi det(\Sigma_i)}} e^{-\frac{1}{2}(x-m_i)^T\Sigma_i^{-1}(x-m_i)}$$

where $m_i = \mathbb{E}[X|Y = i]$ and Σ_i is the covariance matrix for class i. Let $q_0 = P\{Y = 0\}$ and $q_1 = P\{Y = 1\}$ be the a priori probabilities. Determine the Bayes classifier. Characterize the cases when the Bayes decision is linear (i.e., it is obtained by thresholding a linear function of x).

The Bayes classifier is given by

$$g^* = \begin{cases} 1 \text{ if } q_1 f_1(x) > q_0 f_0(x) \\ 0 \text{ otherwise} \end{cases}$$

To determine when $q_1 f_1(x) > q_0 f_0(x)$, take the log of both sides to facilitate an easier equality and reduce to determine when:

$$2\big[log(\frac{q_1}{\sqrt{2\pi det(\Sigma_1)}}) - (x - m_1)^T\Sigma_1^{-1}(x - m_1)\big] > 2\big[log(\frac{q_0}{\sqrt{2\pi det(\Sigma_0)}}) - (x - m_0)^T\Sigma_0^{-1}(x - m_0)\big]$$

Which can be further reduced to:

$$2log(q_1) - log(det\Sigma_1) - (x - m_1)^T\Sigma_1^{-1}(x - m_1) > 2log(q_0) - log(det\Sigma_0) - (x - m_1)^T\Sigma_0^{-1}(x - m_1)$$

We can simplify this expression using the following:

$$r_i^2 = (x - m_i)^T\Sigma_i^{-1}(x - m_i) \textit{ (i.e. the Mahalnoblis distance)}$$

and we get the Bayes classifier is reduced to:

$$g^* = \begin{cases} 1 \text{ if } r_1^2 > r_0^2 + 2log(\frac{q_1}{1-q_1}) + log(\frac{det\Sigma_0}{det\Sigma_1}) \\ 0 \text{ otherwise} \end{cases}$$

When $\Sigma_1 = \Sigma_0 = \Sigma$, the last term is 0:

$$g^* = \begin{cases} 1 \text{ if } r_1^2 > r_0^2 + 2log(\frac{q_1}{1-q_1}) \\ 0 \text{ otherwise} \end{cases}$$

This inequality is linear in x, so the classification rule is linear.

3. Let (X, Y) be a pair of random variables taking values in $X \times \mathbb{R}$ and consider a prediction problem in which one desires to guess the value of Y upon observing X. Suppose that the loss function is $\ell(y, y') = (y - y')^2$. Determine the predictor function $f : X \to \mathbb{R}$ that minimizes the expected loss $E(f(X), Y)$.

The expected loss can be expressed as:

$$\mathbb{E}\ell(y, y') = \int \ell(y, y')p(y|x)dy$$

$$= \int (y - y')^2 p(y|x) dy$$

Where $p(y|x)$ is the conditional distribution of y on x.

To determine the predictor function that minimizes the expected loss, we can take the derivative of the expected loss with respect to y', set to 0 and solve for y':

$$\begin{aligned}
0 &= \frac{\partial}{\partial y'} \int (y - y')^2 p(y|x) dy \\
&= \int \frac{\partial}{\partial y'} \Big[(y - y')^2 p(y|x) \Big] dy \\
&= \int 2(y - y') p(y|x) dy \\
&= 2y' \int p(y|x) dy - 2 \int y p(y|x) dy \\
0 &= y' - \int y p(y|x) dy
\end{aligned}$$

The second term is equivalent the expected value of y at x, thus:

$$y' = \mathbb{E}\big[Y|X = x\big]$$

The predictor function which minimizes the expected loss function is the expected value of Y at $X = x$, in other words the mean of Y at x.

4. Repeat the previous problem but with $\ell(y, y') = |y - y'|$. You may assume that for each $x \in X$, the conditional distribution of Y, given $X = x$, has a density $\phi(y|x)$.

Similar to 3, we can estimate the expected loss in the following way:

$$\begin{aligned}
\mathbb{E}\ell(y, y') &= \int |y - y'| \, \phi(y|x) dy \\
&= \int_{y'}^{-\infty} (y - y') \phi(y|x) dy + \int_{-\infty}^{y'} (y' - y) \phi(y|x) dy
\end{aligned}$$

To find the best prediction function, we minimize the expected loss by taking the derivative, setting to 0 and solving for y':

$$\begin{aligned}
0 &= \frac{\partial}{\partial y'} \int_{y'}^{-\infty} (y - y') \phi(y|x) dy + \frac{\partial}{\partial y'} \int_{-\infty}^{y'} (y' - y) \phi(y|x) dy \\
&= \int_{y'}^{\infty} -\phi(y|x) dy + \int_{-\infty}^{y'} \phi(y|x) dy \\
\int_{y'}^{\infty} \phi(y|x) dy &= \int_{-\infty}^{y'} \phi(y|x) dy
\end{aligned}$$

The above is equivalent to the probability densities:

$$\mathbb{P}(Y \leq y'|x) = \mathbb{P}(Y \geq y'|x)$$

Thus the best predictor function is the y' where these probabilities are equivalent. These are equal at the median of Y at $X = x$.

5. Let $X, X_1, ..., X_n$ be i.i.d. random vectors, uniformly distributed on $[0,1]^d$. Let k be a fixed positive integer and let $X_{(k)}$ denote the k-th nearest neighbor of X among $X_1, ..., X_n$. (We assume $n \geq k$.) Prove that:

$$\lim_{n\to\infty} \|X_{(k)} - X\| = 0 \text{ in probability.}$$

b_d is the unit sphere centered at x, with radius ϵ. The distance of the k nearest neighbors from x can only be greater than the radius of the ball centered at x when there are less than k X_i in the sphere centered at x.

$$\mathbb{P}\{\|X_k(x) - X\| > \epsilon\} = 1 - b_d\epsilon^d$$

which is certainly less than:

$$\mathbb{P}\{\|X_k(x) - X\| > \epsilon\} \leq 1 - \frac{b_d}{2^d}\epsilon^d$$

To simplify we set $c_d = \frac{b_d}{2^d}$:

$$\mathbb{P}\{\|X_k(x) - X\| > \epsilon\} \leq (1 - c_d\epsilon^d)^n$$

Using the inequality $1 + x \leq e^x$:

$$\mathbb{P}\{\|X_k(x) - X\| > \epsilon\} \leq e^{-nc_d\epsilon^d}$$

As n goes to ∞ this term goes to 0 and

$$\|X_k(x) - X\| = 0 \to 1$$

in probability.

6. Show that for any sample size n there exists a distribution of (X, Y) such that $R^* = 0$ but the expected risk of the 1-nearest neighbor classifier is greater than $\frac{1}{4}$.

As described in Theorem 7.1 of *A Probabilistic Theory of Pattern Recognition*, the lower bound for the expected risk of any classifier can be determined by the supremum of the risk for the binary expansion of a uniform random variable $b \in [0, 1)$ which parameterizations any given distribution of (X, Y) as follows:

For any distribution (X, Y), X is defined on the set of positive integers from $\{1, ..., K\}$ where K is an arbitrarily large number to be decided later, such that:

$$p_i = \mathbb{P}(X = i) \begin{cases} \frac{1}{K} \text{ for } i = 1, ..., K \\ 0 \text{ otherwise} \end{cases}$$

A lower bound for the expectation of the error of any given decision rule $g_n(X)$ conditional on the observed distribution of data D_n is $\mathbb{E}[L_n] = R_n(b)$. b is uniformly distributed $[0, 1)$ and acts as a parameter of the distribution of (X, Y) such that it determines the distribution of Y as the binary expansion of b and $b_X = Y$. There exists a b such the risk of the decision rule g_n is at a maximum.

The expected value of the risk $R_n(B)$ must be less than or equal to the maximum risk $R_n(b)$.

$$sup_{b\in[0,1)} R_n(b) \leq \mathbb{E}\{R_n(B)\}$$

The expected value of this random variable, $\mathbb{E}\{R_n(B)\}$ is a lower bound for the expected risk of any given decision rule.

$$
\begin{aligned}
\mathbb{E}\{(R_n(B))\} &= \mathbb{P}\{(g_n(X, D_n)) \neq Y\} \\
&= \mathbb{P}\{(g_n(X, D_n)) \neq B_X\}
\end{aligned}
$$

When g_n is the 1-nearest neighbor rule this becomes

$$
= \mathbb{P}\{(g_n(X, D_n)) \neq B_{X'}\}
$$

Where $B_{X'}$ is the nearest neighbor of B_X when trying to classify X

$$
= \mathbb{P}\{B_{X'} \neq Y\}
$$

$$
\begin{aligned}
&\geq \frac{1}{2}\mathbb{P}\{B_{X'} \neq B_X\}^n \\
&\geq (1 - \frac{1}{K})^n
\end{aligned}
$$

This is $\frac{1}{2}$ as $K \to \infty$. In other words, as the space on which X is defined $\{1, ..., K\}$ grows, the lower bound for the expected risk for any decision rule is $\frac{1}{2} - \epsilon$ where ϵ is a small number.