Elements of statistical learning Ch. 2 exercises

James Chuang

December 20, 2016

Contents

Ex. 2.1 Suppose each of K-classes has an associated target t_k, which is a vector of all zeros, except for a one in the kth position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $\min_k \|t_k - \hat{y}\|$, if the elements of $\hat{y}$ sum to one.

The problem, restated: Show that $\underset{k}{\text{argmin}} \|t_k - \hat{y}\| = \underset{k}{\text{argmax}}(y_k)$ subject to :

$$
\begin{aligned}
& \underset{k}{\text{argmin}} \|t_k - \hat{y}\| \\
&= \underset{k}{\text{argmin}} \|t_k - \hat{y}\|^2 && x \to x^2 \text{ is monotonic} \\
&= \underset{k}{\text{argmin}} \sum_{i=1}^{k} \left(y_i - (t_k)_i\right)^2 && \text{definition of norm, ignoring } \surd \text{ due to argmin} \\
&= \underset{k}{\text{argmin}} \sum_{i=1}^{k} \left(y_i - 2y_i(t_k)_i + (t_k)_i^2\right) \\
&= \underset{k}{\text{argmin}} \sum_{i=1}^{k} \left(-2y_i(t_k)_i + (t_k)_i^2\right) && \sum_{i=1}^{k} y_i^2 \text{ is independent of k} \\
&= \underset{k}{\text{argmin}} \left(-2y_k + 1\right) && \sum_{i=1}^{k} y_i(t_k)_i = y_k, \quad \sum_{i=1}^{k} (t_k)_i^2 = 1 \\
&= \underset{k}{\text{argmin}} \left(-2y_k\right) \\
&= \underset{k}{\text{argmax}} \left(y_k\right)
\end{aligned}
$$

Ex 2.2 Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.

The simulation draws 10 points $p_1, \ldots, p_{10} \in \mathbb{R}^2$ from $N\left(\begin{bmatrix}1\\0\end{bmatrix}, I_2\right)$ and 10 points $q_1, \ldots, q_{10} \in \mathbb{R}^2$ from $N\left(\begin{bmatrix}0\\1\end{bmatrix}, I_2\right)$. These points p_i and q_j we assume to be fixed, and are used as the means of normal distributions with covariance matrix $I_2/5$. The Bayes decision boundary is found by equating the likelihoods of a point being generated from the blue generating function and the orange generating function:

$$
\begin{aligned}
P(\text{blue}) &= P(\text{orange}) \\
\sum_i \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{p}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{p}_i)\right) &= \sum_j \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{q}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{q}_j)\right) \\
\sum_i \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{p}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{p}_i)\right) &= \sum_j \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{q}_j)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{q}_j)\right) \\
\sum_i \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{p}_i)^T\left(\frac{5}{\mathbf{I}_2}\right)(\mathbf{x}-\mathbf{p}_i)\right) &= \sum_j \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{q}_j)^T\left(\frac{5}{\mathbf{I}_2}\right)(\mathbf{x}-\mathbf{q}_j)\right) \\
\sum_i \exp\left(\frac{-5\|\mathbf{p}_i-\mathbf{x}\|^2}{2}\right) &= \sum_j \exp\left(\frac{-5\|\mathbf{q}_j-\mathbf{x}\|^2}{2}\right)
\end{aligned}
$$

Ex 2.3 Derive equation (2.24).

Equation 2.24: Consider N data points uniformly distributed in a p-dimensional unit ball centered at the origin. Suppose we consider a nearest-neighbor estimate at the origin. The median distance from the origin to the closest data point is given by the expression

$$
d(p, N) = \left(1 - \frac{1}{2}^{\frac{1}{N}}\right)^{\frac{1}{p}}
$$

Let $r =$ median distance.

$$
\begin{aligned}
\frac{1}{2} &= P(\text{all } N \text{ points are further than } r \text{ from the origin}) && \text{definition of the median} \\
\frac{1}{2} &= \prod_{i=1}^{N} P(\|x_i\| > r) && \text{each point is assumed to be independent} \\
\frac{1}{2} &= \prod_{i=1}^{N} \left[1 - P\left(\|x_i\| \leq r\right)\right] \\
\frac{1}{2} &= \prod_{i=1}^{N} \left[1 - \frac{Kr^p}{K}\right] && \text{volume of a } p \text{ dimensional hypersphere w/ radius } r \\
\frac{1}{2} &= \prod_{i=1}^{N} \left[1 - r^p\right] \\
\frac{1}{2} &= (1 - r^p)^N \\
1 - r^p &= \left(\frac{1}{2}\right)^{\frac{1}{N}} \\
r^p &= 1 - \left(\frac{1}{2}\right)^{\frac{1}{N}} \\
r &= \left[1 - \left(\frac{1}{2}\right)^{\frac{1}{N}}\right]^{\frac{1}{p}}
\end{aligned}
$$

Ex 2.4 The edge effect problem discussed on page 23 is not peculiar to uniform sampling from bounded domains. Consider inputs drawn from a spherical multinormal distribution $X \sim N(0, \mathbf{I}_p)$. The squared distance from any sample point to the origin has a χ^2_p distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/\|x_0\|$ be an associated unit vector. Let $z_i = a^T x_i$ be the projection of each of the training points on this direction.

Show that the z_i are distributed $N(0, 1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin.

Hence for $p = 10$, a randomly drawn test point is about 3.1 standard deviations from the origin, while all the training points are on average one standard deviation along direction a. So most prediction points see themselves as lying on the edge of the training set.