Machine Learning Lecture Notes

Kede Ma

September 6, 2024

Math Notation

$\mathbb{R}$	the set of real numbers
$\mathbb{R}^N$	N-dimensional real space
M	the number of training samples
N	the number of input features (or data attributes) to represent a training sample
$\mathcal{D}$	the training set that consists of M training samples
x	an n-dimensional input feature vector
$x^{(i)}$	the ith input feature vector in the training set $\mathcal{D}$
$x_j^{(i)}$	the jth feature of the ith input feature vector in $\mathcal{D}$
$y^{(i)}$	the ith output label in $\mathcal{D}$ corresponding to $x^{(i)}$
$\mathcal{Y}$	the set of all output labels
$\lvert\mathcal{Y}\rvert$	the cardinality of $\mathcal{Y}$ (*i.e.*, the number of classes in $\mathcal{Y}$)
$\mathbb{I}[\cdot]$	the indicator function
$\text{sign}(\cdot)$	the sign function
$\sum$	the summation of multiple terms
$\prod$	the product of multiple terms
$\int$	the integration with respect to continuous variables
$\frac{\partial f(z)}{\partial z_j}$	the partial derivative of $f(z)$ w.r.t. z_j, where $z = [z_1, z_2, \ldots, z_N]^T$
$\nabla f(z)$	the derivative of $f(z)$ at z, where $\nabla f(z) = [\frac{\partial f(z)}{\partial z_1}, \ldots, \frac{\partial f(z)}{\partial z_N}]^T$
$\nabla_v f(z)$	the directional derivative of f along the direction v at z
∞	infinity
$\lim_{z \to a} f(z)$	the limit of $f(z)$ as z approaches a
A^T	the transpose of a matrix A
A^{-1}	the inverse of a square matrix A
A^{-T}	the inverse of the transposed A and vice versa, $A^{-T} = (A^{-1})^T = (A^T)^{-1}$
$\text{tr}A$	the trace of a square matrix A
$\lVert A\rVert_F$	the Frobenius norm of a matrix A
$\perp$	perpendicular to

1 Lecture 1

1.1 Derivation of the Bayes Optimal Classifier

Proof. Given a classifier $f \in \mathcal{H}$, where $\mathcal{H}$ denotes the set of candidate classifiers that includes the best one, we first write down the 0-1 loss for a training sample (x, y):

$$\ell(f(x), y) = \mathbb{I}[f(x) \neq y] = 1 - \mathbb{I}[f(x) = y]. \tag{1}$$

The expected predicted error of f is therefore

$$\ell(f) = \mathbb{E}_{x,y}[\ell(f(x), y)], \tag{2}$$

where the expectation is taken w.r.t. the joint distribution $p(x, y)$. Expanding Eq. (2), we have

$$\begin{aligned}
\ell(f) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \ell(f(x), y) \\
&= \sum_{x \in \mathcal{X}} p(x) \left[\sum_{y \in \mathcal{Y}} p(y|x) \ell(f(x), y) \right] \\
&= \mathbb{E}_x \left[\sum_{y \in \mathcal{Y}} p(y|x) \ell(f(x), y) \right].
\end{aligned} \tag{3}$$

As we would like to minimize the expected loss, for each x, we have

$$\begin{aligned}
f_B &= \arg\min_{f \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x) \ell(f(x), y) \\
&= \arg\min_{f \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x)(1 - \mathbb{I}[f(x) = y]) \\
&= \arg\min_{f \in \mathcal{H}} \underbrace{\sum_{y \in \mathcal{Y}} p(y|x)}_{1} - \sum_{y \in \mathcal{Y}} p(y|x) \mathbb{I}[f(x) = y]) \\
&= \arg\min_{f \in \mathcal{H}} - \sum_{y \in \mathcal{Y}} p(y|x) \mathbb{I}[f(x) = y]) \\
&= \arg\max_{f \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x) \mathbb{I}[f(x) = y] \\
&= \arg\max_{f \in \mathcal{H}} p(f(x)|x).
\end{aligned} \tag{4}$$

Thus,

$$f_B(x) = \arg\max_{y \in \mathcal{Y}} p(y|x), \tag{5}$$

and the corresponding expected error rate is

$$\ell = 1 - \mathbb{E}_x \left[\max_{y \in \mathcal{Y}} p(y|x) \right]. \tag{6}$$

In other words, the optimal Bayes decision rule is to choose the class presenting the maximum posterior probability, given the particular observation at hand. Classifiers such as these are called Bayes Optimal Classifiers or Maximum a Posteriori classifiers. □