Statistical Machine Learning

Problem 0 (Naive Bayes)

1. $\hat{y} = argmax_{k\in\{1,2,3\}} g(x_{new}|\mu_k)P(y=k)$
$= argmax_{k\in\{1,2,3\}} \prod_{d=1}^{5} g(x_{new}^{(d)}|\mu_k)P(y=k)$

2. The parameters of the model are taken from the averages of the classified points, i.e.
(a) For each $k \in \{1,2,3\}$, μ_k =average of the x_n for which $y_n = k$.
(b) $P(y=k) = \frac{\text{Number of training points labeled k}}{\text{Number of training points}}$ for $k \in \{1,2,3\}$

3. Yes. The Bayes classifier is optimal when computed from the true distribution. Thus, if we assume that the data source is well approximated by a spherical Gaussian, then the Bayesian classifier will perform close to or exactly optimally.

■

Problem 1 (Maximum Likelihood Estimation)

1. Since $x_1, x_2, ..., x_n$ i.i.d, the conditional density $p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$. In order to find $l(\theta)$, we take the derivative with respect to θ. Because logarithm is monotonically increasing on $\mathbb{R}$, however, we first apply the log, in order to ease calculations. Since the function is concave, if we set the sum of the derivatives of the logs of the function equal to 0, and solve for θ, we will find the maximum likelihood estimator. Thus we solve for μ in the equation $\sum_{i=1}^{n} \nabla_\theta log\, g(x_i|\theta) = 0$.
2.

3.

■

Problem 2 (Bayes-Optimal Classifier)

Outline: We need to show that f_0, the Bayes-optimal classifier, minimizes the conditional risks, $R(f|x)$ for all $x \in \mathbb{R}^d$. Then, we can use the monotonicity of the integral to argue that $R(f)$ is also minimal, and hence the Bayes-optimal classifier minimizes the probability of error (for all integrable functions).

Assume we have some classifier on p, $f_1(x)$, such that $R(f_1|x) \leq R(f_0|x), \forall x \in \mathbb{R}^d$.

Then, $\sum_{y \in \{K\}} L^{0-1}(y, f_1(x))P(y|x) \leq \sum_{y \in \{K\}} L^{0-1}(y, f_0(x))P(y|x)$

which implies $f_1(x) \geq f_0(x) = argmax_{y \in \{K\}} P(y|x)$. However, since f_0 is the max of the $P(y|x)$, f_1 must be exactly equal to f_0, and thus f_0 minimizes the conditional risk $R(f|x)$. Since our choice of x was arbitrary, this holds for any $x \in \mathbb{R}^d$.

As given in the hints, we can now show that since f_0 minimizes $R(f|x)$, $R(f)$ is also minimized, i.e. $R_2(f|x) \geq R(f|x) \Rightarrow \int R_2(f|x)p(x)dx \geq \int R(f|x)p(x)dx$. This is obvious from the conditional risks. If we imagine the conditional risks as the sum of unit impulse functions (of magnitude 0 or 1, representing the loss function) weighted by the conditional probabilities, then the risk $R(f)$ is simply the area under the curve, and therefore if $R(f|x)$ is minimal at every x, then so too, trivially, is its integral over all x. Thus, for $f_0(x) = argmax_{y \in \{K\}} P(y|x)$, the conditional risks $R(f|x)$ are minimized for all x, and therefore so too is the risk $R(f)$, and since the risk is exactly the probability of error, the proof is complete.

■