Machine Learning
Problem Set 5

Ozaner Hansha

December 10, 2020

Question 1

Consider a joint distribution $(\mathbf{x}, y) \sim P(\mathbf{x}, y)$, where $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ and $|\mathcal{Y}| = C$. Recall that the Bayes optimal classifier, which minimizes the $L_{0/1}$ loss for classification of points drawn from this distribution given:

$$h^*(\mathbf{x}) = \arg\max_c p(c \,|\, \mathbf{x}) = c^*$$

Where $p(y \,|\, \mathbf{x})$ is the true conditional distribution of classes given input $\mathbf{x}$. Now consider a different, stochastic, classifier h_r parameterized by another distribution $q(y \,|\, \mathbf{x})$:

$$h_r(\mathbf{x}; q) \sim q(c \,|\, \mathbf{x})$$

Problem: Show that for any distribution q, the risk of h^* is still less than or equal to that of h_r.

Solution: Recall that the risk of a model h is its expected loss which, in this case, is its misclassification rate:

$$R(h) = E_{\mathbf{x},y}[L_{0/1}(h(\mathbf{x}), y)]$$

Let us now calculate the conditional risk of h^*:

$$\begin{aligned}
R(h^*) &= E_{\mathbf{x},y}[L_{0/1}(h^*(\mathbf{x}), y)] && \text{(def. of risk)} \\
&= E_{\mathbf{x},y}[L_{0/1}(\arg\max_{c'} p(c' \,|\, \mathbf{x}), c)] && \text{(def. of } h^*\text{)} \\
&= \int_{\mathbf{x}\in\mathcal{X}} \sum_{c=1}^{C} L_{0/1}(\arg\max_{c'} p(c' \,|\, \mathbf{x}), c) p(\mathbf{x}, c)\, dx && \text{(def. of expectation)} \\
&= \int_{\mathbf{x}\in\mathcal{X}} \sum_{c=1}^{C} L_{0/1}(c^*, c) p(\mathbf{x}, c)\, dx && \text{(def. of } c^*\text{)} \\
R(h^* \,|\, \mathbf{x}) &= \sum_{c=1}^{C} L_{0/1}(c^*, c) p(c \,|\, \mathbf{x}) && \text{(Bayes rules)} \\
&= 1 \cdot p(1 \,|\, \mathbf{x}) + \cdots + 0 \cdot p(c^* \,|\, \mathbf{x}) + \cdots + 1 \cdot p(C \,|\, \mathbf{x}) && (L_{0/1}(c^*, c) = 0 \text{ only when } c = c^*) \\
&= 1 - p(c^* \,|\, \mathbf{x}) && \text{(law of total probability)}
\end{aligned}$$

Now let us compute the conditional risk of h_r for general q:

$$\begin{aligned}
R(h_r; q) &= E_{\mathbf{x},y,q}[L_{0/1}(h_r(\mathbf{x}; q), y)] && \text{(def. of risk)} \\
&= E_{\mathbf{x},y,q}[L_{0/1}(c_r, c)] && \text{(where } c_r \sim q(c \,|\, \mathbf{x})\text{)} \\
&= E_{\mathbf{x},y}\left[\sum_{c'=1}^{C} L_{0/1}(c', c) q(c' \,|\, \mathbf{x})\right] && \text{(def. of expectation)} \\
&= \int_{\mathbf{x} \in \mathcal{X}} \sum_{c=1}^{C} \sum_{c'=1}^{C} L_{0/1}(c', c) q(c' \,|\, \mathbf{x}) p(\mathbf{x}, c)\, dx && \text{(def. of expectation)} \\
R(h_r; q \,|\, \mathbf{x}) &= \sum_{c=1}^{C} \sum_{c'=1}^{C} L_{0/1}(c', c) q(c' \,|\, \mathbf{x}) p(c \,|\, \mathbf{x}) && \text{(Bayes rules)} \\
&= \sum_{c=1}^{C} (1 \cdot q(1 \,|\, \mathbf{x}) + \cdots + 0 \cdot q(c \,|\, \mathbf{x}) + \cdots + 1 \cdot q(C \,|\, \mathbf{x})) p(c \,|\, \mathbf{x}) && (L_{0/1}(c, c) = 0) \\
&= \sum_{c=1}^{C} (1 - q(c \,|\, \mathbf{x})) p(c \,|\, \mathbf{x}) && \text{(law of total probability)} \\
&= \sum_{c=1}^{C} p(c \,|\, \mathbf{x}) - \sum_{c=1}^{C} p(c \,|\, \mathbf{x}) q(c \,|\, \mathbf{x}) \\
&= 1 - \sum_{c=1}^{C} p(c \,|\, \mathbf{x}) q(c \,|\, \mathbf{x}) && \text{(law of total probability)}
\end{aligned}$$

Now let us find the distribution $q(y \,|\, \mathbf{x})$ that would maximize the sum $\sum_{c=1}^{C} p(c \,|\, \mathbf{x}) q(c \,|\, \mathbf{x})$, thus minimizing the risk.

Note that since p is a distribution, each term $p(c \,|\, \mathbf{x})$ can only be lowered or stay unchanged when multiplied by $q(c \,|\, \mathbf{x})$ since $0 \leq q(c \,|\, \mathbf{x}) \leq 1$.

If we had control of both p and q, the sum would be maximized by simply having them equal 1 at an arbitrary class c. However, since we only have control of q, the sum is maximized by simply having $q(c^* \,|\, \mathbf{x}) = 1$, where c^* is the class with the highest posterior probability. The other classes would have to have a 0 probability. We denote this optimal degenerate distribution q^*:

$$\begin{aligned}
R(h_r; q^* \,|\, \mathbf{x}) &= 1 - \sum_{c=1}^{C} p(c \,|\, \mathbf{x}) q^*(c \,|\, \mathbf{x}) && \text{(above)} \\
&= 1 - p(c^* \,|\, \mathbf{x}) && (q^*(c \,|\, \mathbf{x}) = 0 \text{ for all } c \neq c^*)
\end{aligned}$$

Now note two facts, first is that because our chosen q minimizes the risk, all other choices of q either give the same risk or higher. Second is that this risk is identical to that of the Bayes optimal classifier h^*. Putting these together we have:

$$(\forall q \neq q^*) \quad R(h_r; q \,|\, \mathbf{x}) \geq R(h_r; q^* \,|\, \mathbf{x}) = R(h^* \,|\, \mathbf{x})$$

We can put this more succinctly by combining both cases of q:

$$R(h_r; q \,|\, \mathbf{x}) \geq R(h^* \,|\, \mathbf{x})$$

And, of course, multiplying both sides by the marginal probability $p(\mathbf{x})$ nets us:

$$R(h_r; q) \geq R(h^*)$$

Question 2

Problem: Show that the Gaussian naive Bayes classifier has the same form as logistic regression in the two class case.

Solution: Before we begin, let us define $z = P(y = 1)$. This implies that $P(y = 0) = 1 - z$. Now consider the following chain of equalities:

$$
\begin{aligned}
P(y=1\,|\,\mathbf{x}) &= \frac{P(\mathbf{x}\,|\,y=1)P(y=1)}{P(\mathbf{x})} && \text{(Bayes rule)}\\
&= \frac{P(\mathbf{x}\,|\,y=1)P(y=1)}{P(\mathbf{x}\,|\,y=1)P(y=1)+P(\mathbf{x}\,|\,y=0)P(y=0)} && \text{(law of total probability)}\\
&= \frac{1}{1+\frac{P(\mathbf{x}\,|\,y=0)P(y=0)}{P(\mathbf{x}\,|\,y=1)P(y=1)}}\\
&= \frac{1}{1+\exp\left(\ln\frac{P(\mathbf{x}\,|\,y=0)P(y=0)}{P(\mathbf{x}\,|\,y=1)P(y=1)}\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{P(y=0)}{P(y=1)}+\ln\frac{P(\mathbf{x}\,|\,y=0)}{P(\mathbf{x}\,|\,y=1)}\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\ln\frac{P(\mathbf{x}\,|\,y=0)}{P(\mathbf{x}\,|\,y=1)}\right)} && \text{(def. of } z\text{)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\ln\prod_{i=1}^{d}\frac{P(x_i\,|\,y=0)}{P(x_i\,|\,y=1)}\right)} && \text{(conditional independence)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\ln\frac{P(x_i\,|\,y=0)}{P(x_i\,|\,y=1)}\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\ln\frac{(2\pi\sigma_i^2)\exp\left(\frac{-(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{(2\pi\sigma_i^2)\exp\left(\frac{-(x_i-\mu_{i1})^2}{2\sigma_i^2}\right)}\right)} && (p(x_i\,|\,y=c)\sim\mathcal{N}(\mu_{ci},\sigma_i^2))\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\frac{(x_i-\mu_{i1})^2-(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\frac{(x_i^2-2x_i\mu_{i1}-\mu_{i1}^2)-(x_i^2-2x_i\mu_{i0}-\mu_{i0}^2)}{2\sigma_i^2}\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\frac{2x_i(\mu_{i0}-\mu_{i1})+(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}x_i+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)\right)}\\
&= \frac{1}{1+\exp\left(\ln\frac{1-z}{z}+\sum_{i=1}^{d}\left(-w_ix_i+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)\right)} && (\text{define } w_i=-\tfrac{\mu_{i0}-\mu_{i1}}{\sigma_i^2})\\
&= \frac{1}{1+\exp\left(-w_0-\sum_{i=1}^{d}w_ix_i\right)} && (\text{define } w_0=-\ln\tfrac{1-z}{z}-\textstyle\sum_{i=1}^{d}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2})\\
&= \frac{1}{1+\exp\left(-\mathbf{w}\cdot\mathbf{x}-w_0\right)} && \text{(def. of dot product)}
\end{aligned}
$$

And so we have, given the definitions of w_0 and $\mathbf{w}$, that both models have the same form.

Question 3

Problem: Given the same training set, will the rwo models produce the same classifier when trained?

Solution: The modeling assumptions of the naive Bayes classifier, in particular the conditional independence of the input given the class, are not the same as that of the logistic model. The logistic model, for example, does not require that the distribution be conditionally independent in this way. And so when the training set does not conform to this independence assumption, the logistic model will most likely do better (and thus be different) from the Bayes classifier.

Further, Ng & Jordan (2002) showed that the naive Bayes classifier converges faster than than the logistic classifier.

And so, when trained on the same dataset, we might expect the naive Bayes classifier to not match the logistic one.