Massachusetts Institute of Technology

6.867 Machine Learning, Fall 2006

Problem Set 3: Solutions

1. (a) **(5 points)** If A and B are random variables (RVs) with the same probability distribution, then $E[f(A)] = E[f(B)]$. This may be clearer when we write out the corresponding integral:

$$E[f(A)] = \int f(x)p_A(x)dx = \int f(x)p_B(x)dx = E[f(B)]$$

This hold also when A and B are sets of RVs. In particular, $A = \{S_n^{-1}, (\mathbf{x}_1, y_1)\}$, where we would train on S_n^{-1} and test on $(\mathbf{x}_1, y_1)$, has the same distribution as $B = \{S_{n-1}, (\mathbf{x}, y)\}$ where we would train on another set of $n-1$ samples S_{n-1} and test on $(\mathbf{x}, y)$, also sampled from the same distribution.

(b) **(4 points)**This essentially follows from part (a) when we invoke linearity of expectation. First of all, we observe that the argument in part (a) generalizes to:

$$error_1(S_n) = error_2(S_n) = \ldots = error_i(S_n) = \ldots error_n(S_n)$$

So that

$$E[error_{LOOCV}(S_n)] = \frac{1}{n}\sum_{i=1}^{n} E[(y_i - \hat{f}_{-i}(\mathbf{x_i}))^2] = \frac{1}{n}\sum_{i=1}^{n} error_i(S_n) = error_1(S_n)$$

In the first equality of the second equation, we have invoked the Linearity of Expectation Rule to move "E" inside the summation.

(c) **(3 points)** The variances will not be equal: $error_{LOOCV}$ is an estimate based on averaging the error over n trials while $error_1$ is based on a single trial. Recall that if r numbers $v_1, v_2, \ldots, v_r$ are distributed i.i.d., the variance of the sample mean $\bar{v}$ is $1/r$ times the variance of the v_i's (for large r). While the trials in $error_{LOOCV}$ are not independent, $error_{LOOCV}$ will nevertheless have lower variance.

(d) **(4 points)** There are two possible sets of assignments to x_r's such that the training error will be zero: $x_k = y_k$ and $x_k = -y_k$. The former corresponds to f_{keep} and the latter to f_{flip}. In total, there are 2^n possible assignments to x_k's. Thus, the probability of training error being exactly zero is $2/2^n$ or $1/2^{n-1}$.

(e) **(4 points)** Supppose that the classifier chosen after training is $\hat{f} = f_{keep}$ (the argument for the case when f_{flip} is chosen is essentially the same). Given a training error of ϵ, the number of examples with $x_k = y_k$ is $m_1 = n(1 - \frac{\epsilon}{4})$ and number of examples with $x_k = -y_k$ is $m_2 = n\epsilon/4$ where n is the number of examples (note that the previous computation holds for any ϵ). The factor of 4 comes about because each erroneous prediction incurs a penalty of 4 ($= (1-(-1))^2$ or $(-1-(1))^2$)

We claim that if $\epsilon \ll 1/2$, then in any cross-validation step, the chosen classifier $\hat{f}_{-i}$ will be f_{keep}. Here's why: in any cross-validation step, the number of examples of the above categories will change by at most 1, i.e. $|m_{1,-i} - m_1| \leq 1$ and similarly $|m_{2,-i} - m_2| \leq 1$. Clearly, since $m_1 \gg m_2$,

then $m_{1,-i} > m_{2,-i}$ and so that f_{keep} will be chosen in the i-th cross-validation step. In other words, each of the cross-validation classifiers will be the same as the training classifier. Clearly, then training error = cross-validation error.

(f) **(5 points)** When $\epsilon \ll 1/2$, the result of the previous part implies that we can compute the desired quantity solely in terms of the training error. If training error = δ then the classifier is making $n\delta/4$ mistakes (each mistake costs $2^2 = 4$). Suppose that along some dimension i, the training error is $\leq \epsilon$, i.e., the number of mistakes made is at most $\lfloor \frac{n\epsilon}{4} \rfloor$. If the number of mistakes made is k, there are $\binom{n}{k}$ possible ways to arrange them over n examples, each such way having a probability of 2^{1-n} (analogous to results of part (d)). Then, the probability of error being less than ϵ is:

$$p = \sum_{k=0}^{\lfloor \frac{n\epsilon}{4} \rfloor} \binom{n}{k} \frac{1}{2^{n-1}}$$

We require that the probability that at least one dimension lead to error $< \epsilon$ to be at least $1/2$, i.e., the probability that *no* dimension leads to error $< \epsilon$ to be at most $1/2$ i.e.

$$(1-p)^d \leq 1/2$$

This does not include the one dimension that *may be* relevant. Simplifying the expression, we have

$$d \geq \frac{1}{\log_2 \frac{1}{1-p}}$$

2. (a)

$$P(S_n \,|\, \mathcal{J} = \{\ell\}) = \frac{1}{2} \prod_{t=1}^{n} (1 + y_t x_{t\ell})/2 + \frac{1}{2} \prod_{t=1}^{n} (1 - y_t x_{t\ell})/2$$

(b) The problem is that this assigns zero probabilities to getting anything wrong (it doesn't give partial credit). Any technique that assigns a nonzero score to things that are close is fine. The solution we chose for part (c) with $\mathcal{J} = \{\ell\}$ is:

$$P(S_n \,|\, \mathcal{J} = \{\ell\}) = \frac{1}{2} \sum_{\theta \in \{+1,-1\}} \prod_{t=1}^{n} f(y_t \theta, x_{t\ell}),$$

where

$$f(y, x) = \begin{cases} 1 - \varepsilon & \text{if } y = x \\ \varepsilon & \text{if } y \neq x \end{cases}.$$

That is, we assign positive probability to getting it wrong. In general, the marginal likelihood becomes:

$$P(S_n \,|\, \mathcal{J}) = \sum_{\theta \in \{+1,-1\}^{|\mathcal{J}|}} \left(\frac{1}{2}\right)^{|\mathcal{J}|} \prod_{t=1}^{n} \left[\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} f(y_t \theta_j, x_{tj}) \right].$$

(c) As the set becomes larger, the probability will increase (rapidly, at first, and then it will level off). This is because having more features gives us a better chance of choosing a feature that matches (or closely matches) the labels. But the marginal likelihood is a model selection criterion. How is it that we prefer to include more randomly chosen features when the correct model (no dependence between x's and y) has no features at all? The answer lies in the amount of label noise. The correct value of label noise is 0.5 (labels chosen at random without regard to the inputs). Indeed, as you increase the label noise parameter, the marginal likelihood prefers fewer features. Setting the noise level correctly is critical in guiding what aspects of the data the model should try to capture. The noise level would typically be estimated in conjunction with the other parameters.

(d) We would stop adding features when the value of the marginal likelihood starts decreasing.

(e) Now, the likelihood will be virtually unity when the second feature is included and will decrease as more features are included. This is exactly how the model selection criterion should behave. Assuming a low level of label noise is correct in this case.