Suggestions, PS 12, Part 1

15.4. Without loss we can assume that $\mu = 0$, (why?). Show then the correlation

$$\rho(\overline{x}_1^*, \overline{x}_2^*) = \frac{E(\overline{x}_1^*\overline{x}_2^*)}{\sqrt{V(\overline{x}_1^*)V(\overline{x}_2^*)}}\,.$$

Show

$$E(\overline{x}_1^*, \overline{x}_2^*) = \frac{1}{N^2}\sum_{i,i'} E(x_{1i}^*\, x_{2i'}^*) = \frac{N^2}{N^2}E(x_{1i}^*x_{2i'}^*)$$

and

$$E(x_{1i}^*\, x_{2i'}^*) = E(x_{1i}^*x_{2i'}^*)|A)P(A) + E(x_{1i}^*x_{2i'}^*| \sim A)P(\sim A) = \sigma^2 P(A),$$

with A the event that the x_{1i}^* and $x_{2i'}^*$ come from the same sample. Show $P(A) = \frac{1}{N}$, and

$$E(x_{1i}^*\, x_{2i'}^*) = \frac{\sigma^2}{N}. \qquad (*)$$

Thus show $E(\overline{x}_1^*, \overline{x}_2^*) = \frac{\sigma^2}{N}$.

Now you can show

$$V(\overline{x}_1^*) = E\Big[\Big(\frac{1}{N}\sum_i x_{1i}^*\Big)^2\Big] = \frac{1}{N^2}\Big(\sum_{i=i'} E(x_{1i}^*x_{1i'}^*) + \sum_{i\neq i'} E(x_{1i}^*x_{1i'}^*)\Big)$$

$$= \frac{1}{N^2}\Big\{\Big(N\sigma^2) + N(N-1)\cdot E(x_{11}^*x_{12}^*)\Big\}.$$

Why can we just use the prototypes x_{11}^* and x_{12}^* to represent any x_{1i} and $x_{1i'}$ (or, later just the pair $\overline{x}_1^*$ and $\overline{x}_2^*$)?

Now show

$$E(x_{11}^*x_{12}^*) = \frac{\sigma^2}{N},$$

so $E(\overline{x}_{11}^*\overline{x}_{12}^*) = E(x_{11}^*x_{12}^*) = \sigma^2\frac{1}{N}$.

Now you can show

$$V(\overline{x}_1^*) = \frac{1}{N^2}\{(N\sigma^2) + N(N-1)\cdot E(x_{11}^*x_{12}^*)\} = \sigma^2\Big(\frac{2N-1}{N^2}\Big) \quad , \qquad (1)$$

so

$$\rho(\overline{x}_1^*\overline{x}_2^*) = \frac{\sigma^2/N}{\sigma^{2\cdot}\frac{2N-1}{N^2}}.$$

Now $\overline{x}_{bag}$ is the mean of the bootstrap sample averages $\overline{x}_1^*, \overline{x}_2^*, ..., \overline{x}_k^*$. The variance is $E(\overline{x}_i^*) = 0$ (assume):

$$V \quad V(\overline{x}_{bag}) = V\left(\frac{1}{k}\sum_j \overline{x}_j^*\right) = \frac{1}{k^2}\sum_{j,j'} \text{cov}(\overline{x}_j^*, \overset{*}{\overline{x}}_{j'}) = \frac{1}{k^2}(kV(\overline{x}_1^*) + k(k-1)\text{cov}(\overline{x}_1^*, \overline{x}_2^*))$$

$$= \frac{1}{k^2}{}^{\cdot}\left(k\sigma^{2\cdot}\frac{2N-1}{N^2} + k(k-1)\frac{\sigma^2}{N}\cdot\right) = \frac{\sigma^2}{kN^2}((k+1)N-1).$$

Why can you conclude that bagging yields some improvement in variance over the variance (1) of a single bootstrap? Nevertheless, show it does not help much in comparison to the variance $\frac{\sigma^2}{N}$ of the standard estimate $\overline{x}$ of the full sample mean.

12.1. If β,β_0, $\{\xi_i\}_i$ minimize (12.8) then show for a fixed pair $\beta.\beta_0$, the choice of $\{\xi_i\}_i$ must be so that $\sum_i \xi_i$ is minimized (as long as $\xi_i \geq 0$). Therefore show this requires for each i the smallest ξ_i such that $\xi_i \geq 1 - y_i[x_i^T\beta + \beta_0]$ and $\xi_i \geq 0$, i.e.,

$$\xi_i = [1 - y_i(x_i^T\beta + \beta_0)]_+ .$$

Substitute this into the optimization and show (8) is now

$$\min_{\beta,\beta_0,\xi_i} \frac{1}{2}||\beta||^2 + C\sum_i[1 - y_i(x_i^T\beta + \beta_0)]_+ ;$$

show the constraints become irrelevant since they can always be satisfied by some collection of ξ_i no matter what β, β_0 are. Show the optimizations are equivalent with $\lambda = 1/C$.

12.2. Note the discussion in section 12.3.3 (with only a finite number of $h_i(\mathbf{x})$) defines

$$K(\mathbf{x}, \mathbf{x}_i) = \sum_{m=1}^N h_m(\mathbf{x})h_m(\mathbf{x}_i).$$

Thus by (12.28) you can write

$$f(x) = \beta_0 + \sum_{i=1}^N \alpha_i K(\mathbf{x}, \mathbf{x}_i) = \beta_0 + \sum_{m=1}^p\left[\sum_{i=1}^N \alpha_i h_m(\mathbf{x}_i)\right]h_m(\mathbf{x}),$$

getting $\beta_m = \sum_{i=1}^{N} \alpha_i h_m(\mathbf{x}_i)$ by matching the definition of f. Thus show

$$||\beta||^2 = \sum_{m=1}^{p} \sum_{i=1}^{N} \alpha_i h_m(\mathbf{x}_i) \sum_{j=1}^{N} \alpha_j h_m(\mathbf{x}_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = \alpha^T \mathbf{K} \alpha,$$

with $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, proving equivalence of (12.29) and (12.25).

13.3. **Background**: Note that this problem involves two separate inference procedures. (1) The first is the Bayes rule, which assumes that (somehow) we already know the exact joint probability distribution of the feature vector $X = x$ and the class $Y = k$. The Bayes rule is then the best rule for choosing the response k given x (i.e. the one that minimizes the probability of being wrong!).

(2) The second is one-nearest neighbors, i.e., for a test point x finding the closest training point x to it and choosing its (known) class y_i .

We want to express the error rate of (2) in terms of that of (1).

We have K classes, and assume that the underlying probability for class $Y = k$ $(k = 1, ..., K)$ of a feature vector x is given to us as $P(Y = k|x) = p_k(x)$ (i.e., as usual, even with full information x we still have a probability distribution for Y).

To analyze the error rate of method (1), note the Bayes error is the error if we make the best choice given we know the underlying probability model, i.e., if we choose $y = k^*$ where $k^* = \arg\max_{1 \le k \le K} \{p_k(x)\}$, which is the class with the highest underlying probability at test point x. Now given we know (under the Bayes error) that the choice given x will be k^*, the expected error is the probability that the true class y at point x will not be k^*, which is $1 - p_{k^*}(x)$. We define error to be $\text{Err} = \begin{cases} 0 & \text{if } \widehat{f}(x) = y \\ 1 & \text{if } \widehat{f}(x) \neq y \end{cases}$, assuming the correct class for point x is $k = y$. Bayes error is expected error assuming we use the Bayes algorithm, i.e., choose the best $y = k^*$ given the underlying probability model.

On the other hand, if we use one nearest neighbor and the test point $x = x_i$ happens to be one of the training points (as assumed here), then the predicted class of x will be the same as the (true) class $y_i = k$ of x_i, though this not guaranteed to be the class of x, since the class is probabilistically determined even if the feature vector is identical.

Show for the 1-nearest neighbor method above that expected error is

$$\mathbf{E}(\text{Err}) = \sum_{k=1}^{K} \mathbf{E}(\text{Err}|y = k) p_k(x) = \sum_{k=1}^{K} (1 - p_k(x)) p_k(x) \,,$$

using

$$\mathbf{E}(\text{Err}\,|\,y=k) = 1 - \mathbf{P}(\text{Err} \, = \, 0|y=k) \; .$$

Now to try to prove the arithmetic inequality $\sum_{i=1}^{K} p_k(1-p_k) \le 2(1-p_{k^*}) - \frac{K}{K-1}(1-p_{k^*})^2$. First show

$$\sum_{k=1}^{K} p_k - p_k^2 = 1 - \sum_{k=1}^{K} p_k^2 \ge 1 - \sum_{k=1}^{K} p_{k^*} p_k = 1 - p_{k^*}$$

.

Without loss assume $k^* = 1$, i.e., p_1 is largest. Bound $\sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$ above by minimizing $\sum_{k=2}^{K} p_k^2$ subject to $\sum_{k=2}^{K} p_k = 1 - p_1$ (treating p_1 for now as fixed). Using Lagrange multipliers with respect to $p_2, ..., p_k, \lambda$ to minimize $\sum_{k=2}^{K} p_k^2 - \lambda\Big(\sum_{k=2}^{K} p_k - (1-p_1)\Big)$ (with p_1 fixed). From differentiating conclude $2p_k - \lambda = 0 \; (i = 2, ..., k)$, and so all these p_k are equal, i.e.

$$p_k = \frac{1}{K-1}(1-p_1).$$

Thus show the maximum value of $\sum_{k=1}^{K} p_k(1-p_k)$ is

$$p_1(1-p_1) + \sum_{k=2}^{K} p_k(1-p_k) = p_1(1-p_1) + \sum_{k=2}^{K} \frac{1-p_1}{K-1}\Big(\frac{1-p_1}{K-1}\Big)$$

$$= p_1(1-p_1) + (K-1)\frac{1-p_1}{K-1}\Big(1 - \frac{1-p_1}{K-1}\Big)$$

$$= 2(1-p_1) - (1-p_1)^2\left(\frac{K}{K-1}\right).$$

Technically you should also check the boundary values, e.g. where $p_2 = 1 - p_1$ and all other $p_i (i \ne 1, 2)$ are zero, but show this is easy and the maximum is as above.

Letting the Bayes error rate be $E^* = 1 - p_{k^*}(x) = 1 - p_1(x)$, conclude the 1-nearest neighbor error rate is bounded by $E^*\Big(2 - E^* \frac{K}{K-1}\Big)$. Conclude
that if the data size gets large, 1-nearest neighbor almost always finds a training point x_1 that is very close to the test point x (even if they are not identical as above). Assuming that the underlying Bayes distribution is continuous in x, show this will mean that the distribution $p_k(x)$ (as a function of k) is almost identical to $p_k(x_i)$, where x_i is the training point closest to x. Show this allows us to use the above bound with increasing

accuracy as the sample size gets large. The statement in the problem about L^1 convergence is not clear and can be omitted.