Homework 2

Q1. Derive expression [1, (10.12) on p.344] for the update parameter in AdaBoost.

S1. Note that we need to minimize the expression

$$\left(e^{\beta} - e^{-\beta}\right) \sum_{i=1}^{N} w_i^m I(y_i \neq G_m(x_i)) + e^{-\beta} \sum_{i=1}^{N} w_i^m$$

(cf. [1, (10.11) p. 344]) with respect to β. For this, we differentiate with respect to β, determine the zeros and get

$$e^{2\beta} = \frac{\sum_{i=1}^{N} w_i^m - \sum_{i=1}^{N} w_i^m I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i^m I(y_i \neq G_m(x_i))}.$$

Using the definition of the error at m-th stage (cf. [1, p. 339]) gives

$$e^{2\beta} = \frac{1}{\mathrm{err}_m} - 1 = \frac{1 - \mathrm{err}_m}{\mathrm{err}_m}.$$

This easily gives the desired result.

Q2. Multiclass exponential loss (see also [2]). For a K-class classification problem, consider the coding $Y = (Y_1, ..., Y_K)^T$ with

$$Y_k := \begin{cases} 1 & \text{if } G = \mathcal{G}_k \\ -\frac{1}{K-1}, & \text{otherwise.} \end{cases}$$

Let $f = (f_1, ..., f_K)^T$ with $\sum_{k=1}^{K} f_k(x) = 0$, and define

$$L(Y, f) = \exp\left(-\frac{1}{K} Y^T f\right). \tag{1}$$

a) Using Lagrange multipliers, derive the population minimizer f^* of $E(Y, f)$, subject to the zero-sum constraint, and relate these to the class probabilities.

b) Show that a multiclass boosting using this loss function leads to a reweighting algorithm similar to Adaboost, as in [1, Section 10.4].

S2a. We need to determine

$$f^*(x) = (f_1^*(x), ..., f_K^*(x))^T := \arg\min_{f(x)} \mathbb{E}_{Y|x}\left[L(Y, f(x))\right] = \arg\min_{f(x)} \mathbb{E}_{Y|x}\left[\exp\left(-\frac{1}{K} Y^T f(x)\right)\right]$$

provided $\sum_{k=1}^{K} f_k = 0$. Note that

$$\mathbb{E}_{Y|x}\left[\exp\left(-\frac{1}{K} Y^T f(x)\right)\right] = \sum_{l=1}^{K} \exp\left(\sum_{i \neq l} \frac{f_i(x)}{K(K-1)} - \frac{f_l(x)}{K}\right) \Pr\left(G = \mathcal{G}_l | x\right).$$

Using the constraint $\sum_k f_k = 0$ and writing λ for the Lagrange multiplier, this yields the Lagrangian objective function

$$\sum_{i=1}^{K} \exp\left(-\frac{f_i(x)}{K-1}\right) \Pr(G = \mathcal{G}_i | x) - \lambda \sum_{i=1}^{K} f_i(x).$$

Taking derivatives with respect to f_k, λ and equating to 0, we get for $k = 1, ..., K$

$$f_k^*(x) = (K-1) \log \Pr(G = \mathcal{G}_k | x) - \frac{K-1}{K} \sum_{k'=1}^{K} \log \Pr(G = \mathcal{G}_{k'} | x).$$

This gives

$$\Pr\left(G=\mathcal{G}_k|x\right)=\left\{\prod_{k'=1}^{K}\Pr(G=\mathcal{G}_{k'}|x)\right\}^{1/k} e^{\frac{f_k^*(x)}{K-1}}.$$

Summing both sites from $k=1$ to $k=K$ and using the previous equation again, we get

$$\Pr(G=\mathcal{G}_k|x)=\frac{e^{\frac{1}{K-1}}}{\sum_{k=1}^{K} e^{\frac{1}{K-1}f_k^*(x)}}.$$

S2b. Define the following algorithm

Algorithm 0.1 (SAMME) *cf. [2, p. 351]*

1. *Initialize the observation weights* $w_i = 1/n$, $i=1,...,n$.
2. *For* $m=1$ *to* M*:*
   (a) *Fit a classifier* $G_m(x)$ *to the training data using weights* w_i.
   (b) *Compute*
   $$err_m=\frac{\sum_{i=1}^{n} w_i I\left(y_i\neq G_m(x_i)\right)}{\sum_{i=1}^{N} w_i}$$
   (c) *Compute*
   $$\alpha_m=\log\frac{1-err_m}{err_m}+\log\left(K-1\right)$$
   (d) *Set*
   $$w_i\leftarrow w_i\cdot\exp\left(\alpha_m I(y_i\neq G_m(x_i))\right)$$
   for $i=1,...,n$.
3. *Output* $G(x)=\arg\max_k\sum_{m=1}^{M}\alpha_m I(G_m(x)=k)$

Note the similarity between SAMME and AdaBoost: they coincide when $K=2$ and if $K>2$ the difference is the term $\log(K-1)$ in step (c) of SAMME.

Similar to the case $K=2$ (covered in [1, pp. 343-344]), it can be showed that SAMME is equivalent to fitting a stagewise additive model using the multi-class exponential loss function (1). See [2, pp.352-353] for details.

Q3. Derive the variance formula [1, (15.1), p. 588]. This appears to fail if ρ is negative; diagnose the problem in this case.

S3. Let $X_1,...,X_B$ be a collection of B identically distributed (but not necessarily independent) random variables such that $X_1\sim\mathcal{N}(\mu,\sigma^2)$. Denote by ρ the the positive, pairwise correlation coefficient. By definition

$$0<\rho:=\frac{E\left[(X_i-\mu)(X_j-\mu)\right]}{\sigma^2}\quad\Longrightarrow\quad E(X_iX_j)=\rho\sigma^2+m^2.$$

We then find

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{B}\sum_{i=1}^{B}X_i\right)&=\frac{1}{B^2}\left\{E\left[\left(\sum_{i=1}^{B}X_i\right)^2\right]-E\left[\sum_{i=1}^{B}X_i\right]^2\right\}\\&=\rho\sigma^2+\frac{1-\rho}{B}\sigma^2.\end{aligned}$$

Q4. Suppose x_i, $i=1,...,N$ are iid (μ,σ^2). Let $\bar{x}_1^*$ and $\bar{x}_2^*$ be two bootstrap realizations of the sample mean. Show that the sampling correlation $\mathrm{corr}(\bar{x}_1^*,\bar{x}_2^*)=\frac{n}{2n-1}\approx 50\%$. Along the way, derive $\mathrm{var}(\bar{x}_1^*)$ and the variance of the bagged mean $\bar{x}_{bag}$. Here $\bar{x}$ is a *linear* statistic; bagging produces no reduction in variance for linear statistics.

S4. Let $x_1, ..., x_n$ be iid Gaussian with $x_1 \sim \mathcal{N}(\mu, \sigma^2)$. For $i = 1, 2$ let $\bar{x}_i^* := \frac{1}{n}\sum_{j=1}^n x_{ij}$, where x_{ij} are randomly drawn with replacement from $\{x_1, ..., x_n\}$. Note

$$\mathrm{Cov}(x_{ij}, x_{lk}) = \sigma^2 \mathrm{Pr}\left(x_{ij} = x_{lk}\right) = \frac{\sigma^2}{n} \Rightarrow \mathrm{Cov}(\bar{x}_1^*, \bar{x}_2^*) = \frac{\sigma^2}{n}$$

and

$$\mathrm{Var}(\bar{x}_1^*) = \frac{1}{n^2}\sum_{i,j=1}^n \mathrm{Cov}(x_{1i}, x_{1j}) = \frac{2n-1}{n^2}\sigma^2.$$

Together, the previous two lines give

$$\mathrm{corr}(\bar{x}_1^*, \bar{x}_2^*) = \frac{n}{2n-1}.$$

We also get

$$\mathrm{Var}(\bar{x}_{bag}) = \mathrm{Var}\left(\frac{1}{2}(\bar{x}_1^* + \bar{x}_2^*)\right) = \frac{3n-1}{2n^2}\sigma^2.$$

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer, New York, 2009.

[2] Ji Zhu, Hui Zou, Saharon Rosset and Trevor Hastie. Multi-class adaboost *Statistics and Interface*, 2:349-360, 2009.