

# General PECOC analysis

November 16, 2008

## 1 Reduction

Let  $h : \mathcal{X} \rightarrow \mathbb{R}^d$  be a possibly randomized target function we wish to learn. In multiclass problems, we assume that  $h(x) \in \{e_1, \dots, e_d\}$ , so  $\mathbb{E}h(x) \in \Delta^d$ ; in multilabel problems, we assume that  $h(x) \in \{0, 1\}^d$ , so  $\mathbb{E}h(x) \in [0, 1]^d$ . We are interested in reducing the problem of learning the vector-valued function  $h$  to the problem of learning a collection of scalar-valued functions.

Let  $C = [c_1 | \dots | c_m]^\top \in \mathbb{R}^{m \times d}$  be our coding matrix (so each  $c_i \in \mathbb{R}^d$ ), for some  $m \leq d$ . We create  $m$  regression problems; the label of  $x \in \mathcal{X}$  in the  $i$ th problem is  $\langle c_i, h(x) \rangle$ . The regressor  $b_i^*$  with least mean-squared-error for the  $i$ th problem is defined by  $b_i^*(x) = \mathbb{E}\langle c_i, h(x) \rangle = \langle c_i, \mathbb{E}h(x) \rangle$ . Let  $b^* : \mathcal{X} \rightarrow \mathbb{R}^m$  be the vector-valued regressor defined by  $b^*(x) = [b_1^*(x), \dots, b_m^*(x)]^\top = C\mathbb{E}h(x)$ .

## 2 Recovery for invertible $C$

If  $C \in [-1, +1]^{d \times d}$  is invertible, then we can recover  $\mathbb{E}h(x)$  from  $b^*(x)$  by taking

$$C^{-1}b^*(x) = C^{-1}C\mathbb{E}h(x) = \mathbb{E}h(x).$$

Given regressors  $b_1, \dots, b_d$ , we are interested in the performance of the vector-valued regressor  $C^{-1}b$  defined by  $C^{-1}b(x) = C^{-1}[b_1(x), \dots, b_d(x)]^\top$ .

### 2.1 Generic regret statement for invertible $C$

Consider any regressor  $b : \mathcal{X} \rightarrow \mathbb{R}^d$ . Then the regret of  $C^{-1}b$  is

$$\mathbb{E}_x \|C^{-1}b(x) - \mathbb{E}h(x)\|_2^2 = \mathbb{E}_x \|C^{-1}(b(x) - b^*(x))\|_2^2.$$

We would like to bound this quantity in terms of the average (scaled) regret of the  $b_i$ , i.e. in terms of

$$\bar{r} = \frac{1}{d} \sum_{i=1}^d r_i,$$

where

$$r_i = \mathbb{E}_x \left( \frac{b_i(x) - b_i^*(x)}{B_i} \right)^2 \quad \text{and} \quad B_i = \max_x \langle c_i, h(x) \rangle - \min_x \langle c_i, h(x) \rangle.$$

### 2.2 One-against-all reduction

If  $C = C^{-1} = I_d$  is the identity matrix, then the regret of  $C^{-1}b$  is

$$\mathbb{E}_x \|C^{-1}(b(x) - b^*(x))\|_2^2 = \mathbb{E}_x \|b(x) - b^*(x)\|_2^2.$$

If  $B_i \leq B$  for all  $i$ , then the regret is bounded by  $dB^2\bar{r}$ . If  $\|h(x)\|_\infty \leq 1$  (e.g. in multiclass and multilabel problems), then  $B \leq 1$ .

### 2.3 Hadamard reduction

Let  $C \in \{\pm 1\}^{d \times d}$  be the Hadamard matrix. The singular values of  $C$  are all  $\sqrt{d}$ , so  $\|C^{-1}\|_2^2 = 1/d$ . Therefore the regret of  $C^{-1}b$  is

$$\mathbb{E}_x \|C^{-1}(b(x) - b^*(x))\|_2^2 \leq \|C^{-1}\|_2^2 \mathbb{E}_x \|b(x) - b^*(x)\|_2^2 = \frac{1}{d} \mathbb{E}_x \|b(x) - b^*(x)\|_2^2.$$

If  $B_i \leq B$  for all  $i$ , then the regret is bounded by  $B^2 \bar{r}$ . If  $\|h(x)\|_1 \leq k$  (e.g. in  $k$ -sparse multilabel problems), then  $B \leq 2k$ .

### 2.4 Ridge regression

Suppose  $\mathcal{X} = \mathbb{R}^p$  and we use ridge regression to learn the  $b_i = (\widehat{\mathbb{E}}[xx^\top] + \lambda I_p)^{-1} \widehat{\mathbb{E}}[x \langle h(x), c_i \rangle]$ . Then  $b = (\widehat{\mathbb{E}}[xx^\top] + \lambda I_p)^{-1} \widehat{\mathbb{E}}[xh(x)^\top] C^\top$ . The prediction on a new point  $x_{\text{new}}$  is

$$\begin{aligned} C^{-1} b^\top x_{\text{new}} &= C^{-1} C \widehat{\mathbb{E}}[xh(x)^\top]^\top (\widehat{\mathbb{E}}[xx^\top] + \lambda I_p)^{-1} x_{\text{new}} \\ &= \widehat{\mathbb{E}}[xh(x)^\top]^\top (\widehat{\mathbb{E}}[xx^\top] + \lambda I_p)^{-1} x_{\text{new}}. \end{aligned}$$

Thus the predictor is the same as the one learned using ridge regression but without the reduction.