CSC2515 — Assignment #1 Answers

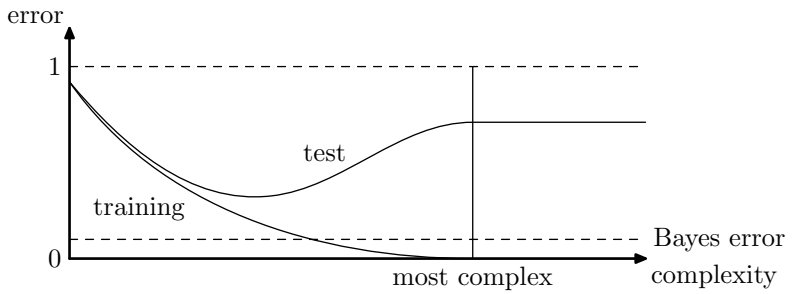
Behdad Esfahbod*
993505827

October 19, 2004

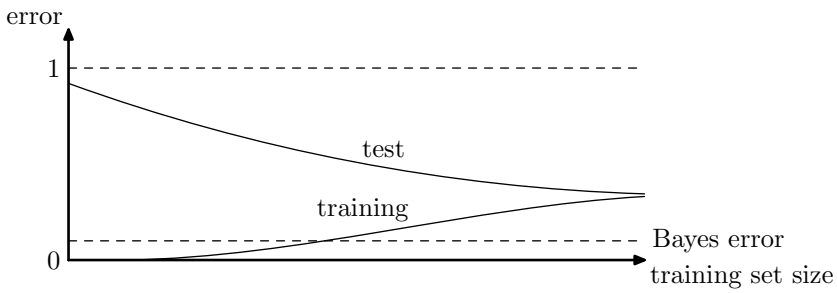
1 Training/Testing Error Curves (1.5%)

The crossing pair of lines are exactly the ones as in graphs below. In other words, training set error crosses Bayes error, but test set error does not.

- Fixed training set size:



- Fixed model complexity:



*Not registered in the course yet. Already talked to Sam and will do late-registration after administration problem is resolved.

2 Learning Random Boolean Functions (2.5%)

- **a:** Let $\mathcal{N}$ be the multiset of training cases.

$$a = E(\sum_{x \in \mathcal{N}} A_x) = \sum_{x \in \mathcal{N}} E(A_x)$$

where A_x is a random variable that is zero if $x \notin \mathcal{N}$ and 1 otherwise.

$$A_x = 1 - p(x \notin \mathcal{N}) = 1 - \prod_{x_0 \in \mathcal{N}} p(x \neq x_0) = 1 - (\frac{2^k - 1}{2^k})^N$$

Putting together:

$$a = \sum_{x \in \mathcal{N}} E(A_x) = 2^k \left(1 - (\frac{2^k - 1}{2^k})^N\right) = \frac{2^k - (2^k - 1)^N}{2^{k(N-1)}}$$

- **b:** Let $\mathcal{M}$ be the multiset of test cases.

$$b = E(\sum_{x \in \mathcal{M}} B_x) = \sum_{x \in \mathcal{M}} E(B_x)$$

where B_x is a random variable that is zero if $x \notin \mathcal{N}$ and 1 otherwise.

$$B_x = p(x \in \mathcal{N}) = \frac{a}{2^k}$$

Putting together:

$$b = \sum_{x \in \mathcal{M}} E(B_x) = \frac{Ma}{2^k}$$

- **Lowest error rate:**

$$\frac{0 * b + \frac{1}{2} * (M - b)}{M} = \frac{M - b}{2M}$$

Algorithm: If test case has been in training cases, return the output to one such training case; return 0 otherwise.

Argument: For seen cases, error is zero. For the rest, they are independently and uniformly distributed among classes, so no difference what to decide.

- **Generalization:** Yes. Because the error rate

$$\frac{M - b}{2M} = 1 - \frac{a}{2^k}$$

is independent of the test set.

3 Class-Conditional Gaussians (3%)

- $$\begin{aligned} p(y=k|\mathbf{x}) &= \frac{p(\mathbf{x}|y=k)p(y=k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|y=k)\alpha_k}{\sum_{k=1}^{K} p(\mathbf{x}|y=k)\alpha_k} \\ &= \frac{\alpha_k \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i-\mu_{ki})^2\right\}}{\sum_{k=1}^{K} \alpha_k \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i-\mu_{ki})^2\right\}\alpha_k} \end{aligned}$$

- $$\begin{aligned} \ell(\theta; D) &= \log p(y^1, x^1, y^2, x^2, \ldots, y^M, x^M|\theta) \\ &= \log \prod_{j=1}^{M} p(y^j, x^j|\theta) = \sum_{j=1}^{M} \log p(y^j, x^j|\theta) = \sum_{j=1}^{M} \left(\log p(x^j|y^j, \theta) + \log p(y^j)\right) \\ &= \sum_{j=1}^{M} \left[-\frac{1}{2}\sum_{i=1}^{D} \log(2\pi\sigma_i^2) - \sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i^j - \mu_{y_j i})^2 + \log \alpha_{y^j}\right] \end{aligned}$$

- $$\begin{aligned} \frac{\partial \ell(\cdot)}{\partial \mu_{ki}} &= \frac{1}{\sigma_i^2} \sum_{j:\, y^j=k} (\mu_k - x_i^j) \\ \frac{\partial \ell(\cdot)}{\partial \log \sigma_i^2} &= \sum_{j=1}^{M} \frac{\partial}{\partial \log \sigma_i^2}\left[-\frac{1}{2}\log \sigma_i^2 - \frac{1}{2} e^{-\log \sigma_i^2}(x_i^j - \mu_{y_j i})^2\right] \\ &= -\frac{M}{2} + \frac{1}{2\sigma_i^2}\sum_{j=1}^{M} (x_i^j - \mu_{y_j i})^2 \\ \frac{\partial \ell(\cdot)}{\partial \alpha_k} &= \sum_{j:\, y^j=k} \frac{1}{\alpha_k} \end{aligned}$$

- $$\begin{aligned} \sum_{j:\, y^j=k} (\mu_k - x_i^j) = 0 \quad &\Rightarrow \quad \mu_k = \frac{\sum_{j:\, y^j=k} x_i^j}{\sum_{j:\, y^j=k} 1} \\ -\frac{M}{2} + \frac{1}{2\sigma_i^2}\sum_{j=1}^{M} (x_i^j - \mu_{y_j i})^2 = 0 \quad &\Rightarrow \quad \sigma_i^2 = \frac{1}{M}\sum_{j=1}^{M} (x_i^j - \mu_{y^j i})^2 \end{aligned}$$

  Using Lagrange Multipliers for class priors to sum them up to one:

$$L(\alpha, \lambda) = \ell(\cdot) + \lambda c(\alpha)$$

$$c(\alpha) = 1 - \sum_{k=1}^{K} \alpha_k$$

  Means:

$$\frac{\partial L(\cdot)}{\partial \alpha} = 0 \quad \Rightarrow \quad \lambda + \sum_{j:\, y^j=k} \frac{1}{\alpha_k} = 0$$

$$\frac{\partial L(\cdot)}{\partial \lambda} = 0 \quad \Rightarrow \quad 1 - \sum_{k=1}^{K} \alpha_k = 0$$

Solving for λ first and σ_k next gives:

$$\begin{aligned} \lambda &= M \\ \alpha_k &= \frac{\sum_{j:\, y^j = k} 1}{M} \end{aligned}$$

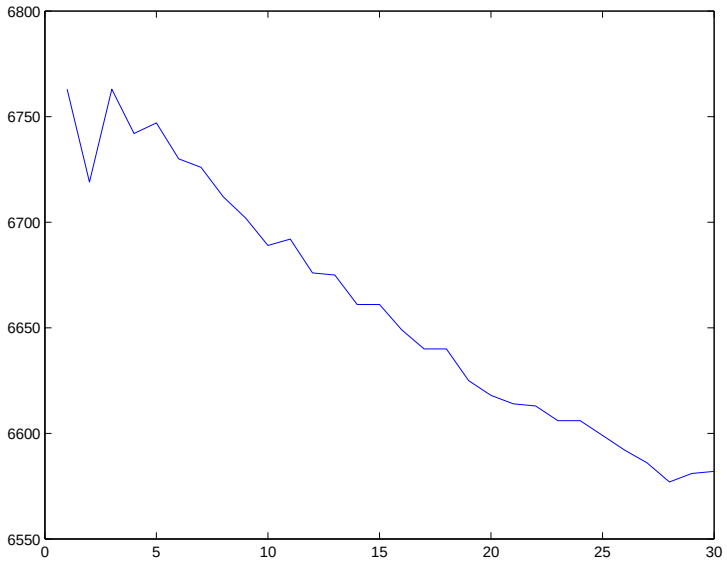
and observe that fortunately α_k's are all positive and less than one.

4 Handwritten Digit Classification (11%)

The images are in row-major, top to bottom, left to right.

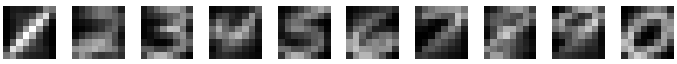
4.1 K-NN Classifier

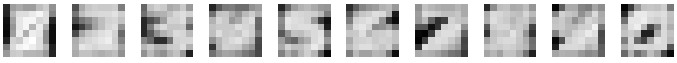
I broke ties by getting the first item, i.e. the one with smallest index. This way $k = 1$ turned out to be the best k as can be seen in the graph below:



4.2 Conditional Gaussian Classifier Training

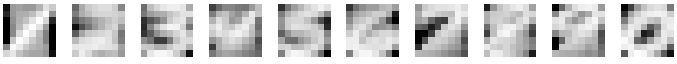
For the means, I took black for zero, and white for maximum value over all means, using linear grayscale in between. For log of covariances, mapped black to the minimum over all values of the diagonals and white to the maximum of them, again using linear grayscale in between. The image of covariances is a lot less *blurred* if I draw covariance diagonals themselves instead of the log of them and map colors linearly.





4.3 Naïve Bayes Classifier Training

Black represents the minimum over all values of the logs and white represents the maximum of them, using linear grayscale in between.



4.4 Performance Evaluation

	Gaussian-conditional	Naïve Bayes
training set	-63.7881	-12.7559
test set	-62.3942	-12.7017

Table 1: Average conditional log likelihood

	K-NN	Gaussian-conditional	Naïve Bayes
1	1 (0.25%)	0 (0.00%)	162 (23.14%)
2	11 (2.75%)	5 (1.25%)	174 (24.86%)
3	21 (5.25%)	13 (3.25%)	141 (20.14%)
4	14 (3.50%)	11 (2.75%)	147 (21.00%)
5	19 (4.75%)	17 (4.25%)	178 (25.43%)
6	7 (1.75%)	9 (2.25%)	93 (13.29%)
7	13 (3.25%)	14 (3.50%)	157 (22.43%)
8	26 (6.50%)	21 (5.25%)	225 (32.14%)
9	11 (2.75%)	15 (3.75%)	227 (32.43%)
0	2 (0.50%)	4 (1.00%)	77 (11.00%)
total	125 (3.125%)	109 (2.725%)	1581 (22.59%)

Table 2: Error cases on test sets

	K-NN	Gaussian-conditional	Naïve Bayes
1	1 (0.14%)	3 (0.43%)	87 (21.75%)
2	25 (3.57%)	12 (1.71%)	106 (26.50%)
3	26 (3.71%)	11 (1.57%)	91 (22.75%)
4	31 (4.43%)	19 (2.71%)	85 (21.25%)
5	22 (3.14%)	13 (1.86%)	110 (27.50%)
6	9 (1.29%)	5 (0.71%)	60 (15.00%)
7	23 (3.29%)	12 (1.71%)	89 (22.25%)
8	57 (8.14%)	27 (3.86%)	122 (30.50%)
9	34 (4.86%)	22 (3.14%)	134 (33.50%)
0	9 (1.29%)	6 (0.86%)	59 (14.75%)
total	237 (3.39%)	130 (1.86%)	943 (23.58%)

Table 3: Error cases on training sets