Extrapolating prediction error for 'extreme' multi-class classification

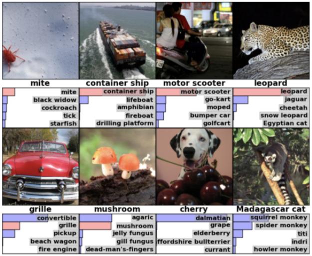
Charles Zheng

Stanford University

February 13, 2017

(Joint work with Rakesh Achanta and Yuval Benjamini.)

Multi-class classification

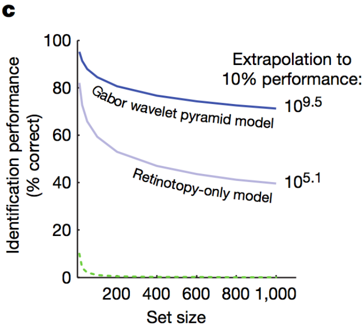


from Krizhevsky et al. 2012

- MNIST digit recognition: 10 categories
- Human motion database: 51 categories
- ImageNet: 22,000 categories
- Wikipedia: 325,000 categories

Accuracy vs. number of classes

Kay (2008) image identification task in functional MRI.



- Question: how does the accuracy scale with the number of classes?

Setup

1. Population of categories $\pi(y)$

2. Subsample k labels, $y_1, \ldots, y_k$

Setup

1. Population of categories $\pi(y)$

2. Subsample k labels, $y_1, \ldots, y_k$

3. Collect training and test data $x_i^{(j)}$ for labels $\{y_1, \ldots, y_k\}$.

Setup

1. Population of categories $\pi(y)$

2. Subsample k labels, $y_1, \ldots, y_k$

3. Collect training and test data $x_i^{(j)}$ for labels $\{y_1, \ldots, y_k\}$.
4. Train a classifier and compute test error.

Setup

1. Population of categories $\pi(y)$

2. Subsample k labels, $y_1, \ldots, y_k$

3. Collect training and test data $x_i^{(j)}$ for labels $\{y_1, \ldots, y_k\}$.
4. Train a classifier and compute test error.

Can we analyze how error depends on k?

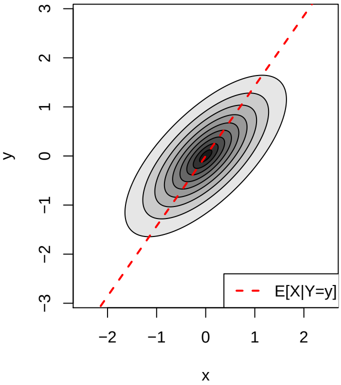
Toy example

$Y_1, \ldots, Y_k \overset{iid}{\sim} N(0, 1)$;

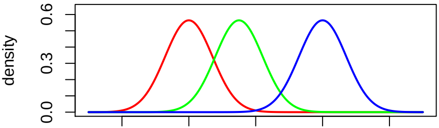
Toy example

$Y_1, \ldots, Y_k \overset{iid}{\sim} N(0,1)$;

$X|Y \sim N(\rho Y, 1-\rho^2)$ i.e. $(Y, X) \sim N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$.

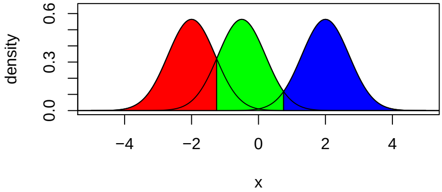


Toy example



- Suppose $k = 3$, and we draw Y_1, Y_2, Y_3.
- The *Bayes rule* is the optimal classifier and depends on knowing the true densities:
$$\hat{y}(x) = \text{argmax}_{y_i} p(x|y_i)$$
- The *Bayes Risk*, which is the misclassification rate of the optimal classifier.

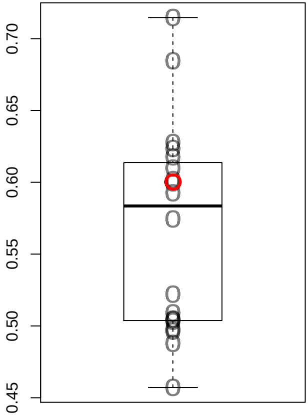
Toy example

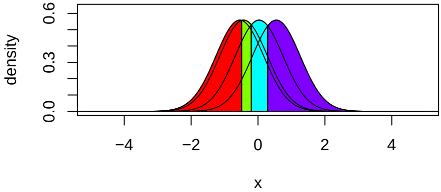


- The *Bayes Risk* is the expected test error of the Bayes rule,

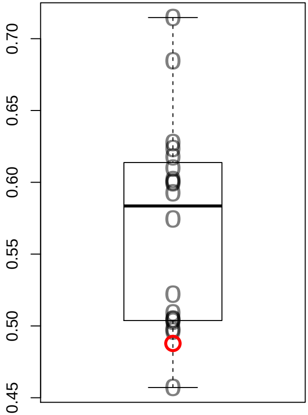
$$\frac{1}{k}\sum_{i=1}^{k}\Pr[\hat{y}(x) \neq Y | Y = y_i]$$

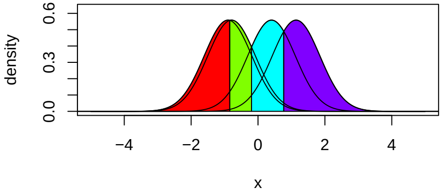
Toy example



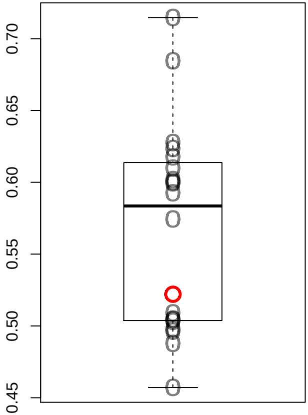


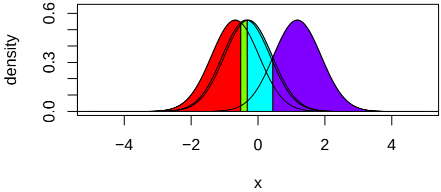
Toy example



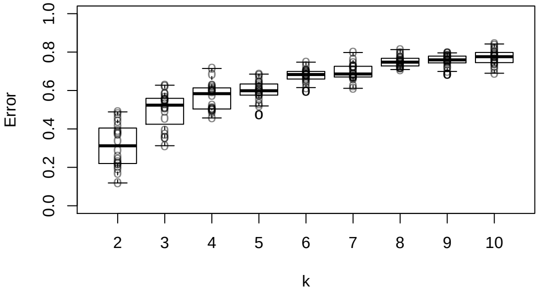


Toy example

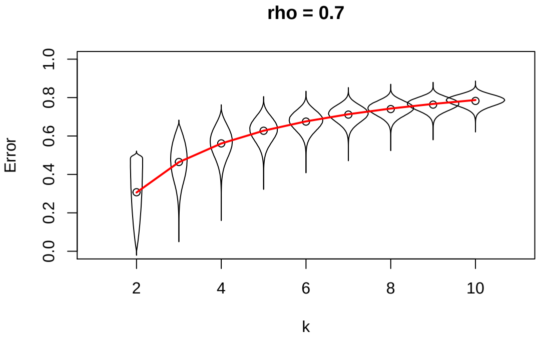




Toy example



Toy example



Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. Flip a coin to choose a class $i \in \{1, 2\}$.
3. Draw X from the ith class, $X \sim p(x|y_i)$.

Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. WLOG assume the true class is $i = 1$.
3. Draw X from the first class, $X \sim p(x|y_1)$.

Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. WLOG assume the true class is $i = 1$.
3. Draw X from the first class, $X \sim p(x|y_1)$.
4. Correct classification if $p(X|y_1) > p(X|y_2)$.

Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. WLOG assume the true class is $i = 1$.
3. Draw X from the first class, $X \sim p(x|y_1)$.

Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. WLOG assume the true class is $i = 1$.
3. Draw X from the first class, $X \sim p(x|y_1)$.
4. Correct classification if $|X - \rho y_1| < |X - \rho y_2|$.

Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. WLOG assume the true class is $i = 1$.
3. Draw X from the first class, $X \sim p(x|y_1)$.
4. Correct classification if $|X - \rho y_1| < |X - \rho y_2|$.

The Bayes risk for labels y_1, y_2 is

$$\text{Risk}(y_1, y_2) = \Pr[|X - \rho y_1| < |X - \rho y_2|].$$

Considering the $k = 2$ case

How is the risk in $k = 2$ case defined?

1. Draw $Y_1, Y_2 \sim N(0, 1)$.
2. WLOG assume the true class is $i = 1$.
3. Draw X from the first class, $X \sim p(x|y_1)$.
4. Correct classification if $|X - \rho y_1| < |X - \rho y_2|$.

The *average* Bayes Risk for $k = 2$ is

$$\text{AvRisk}_2 = \mathbf{E}[\text{Risk}(Y_1, Y_2)] = \Pr[|X - \rho Y_1| < |X - \rho Y_2|].$$

for $X \sim N(\rho Y_1, 1 - \rho^2)$.

Considering the $k = 2$ case

The main theoretical trick we use is to change the order of conditioning.