Extrapolating prediction error for ‘extreme’ multi-class classification

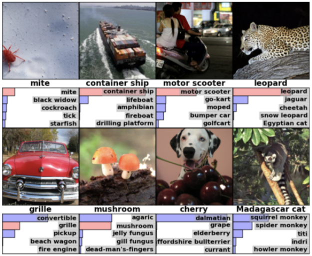
Charles Zheng

Stanford University

February 19, 2017

(Joint work with Rakesh Achanta and Yuval Benjamini.)

Multi-class classification



from Krizhevsky et al. 2012

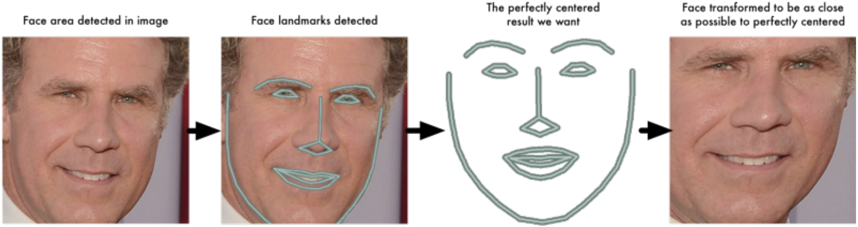
- MNIST digit recognition: 10 categories
- Human motion database: 51 categories
- ImageNet: 22,000 categories
- Wikipedia: 325,000 categories

Facial recognition

- Used to tag images in software, security

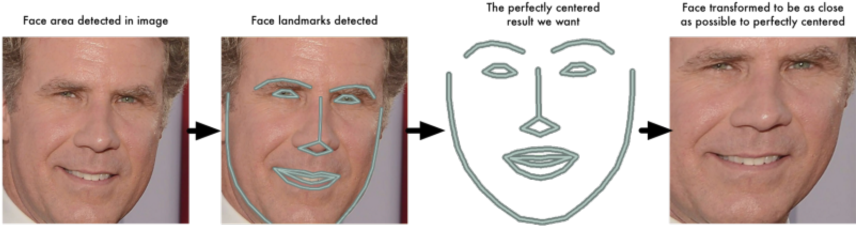
Facial recognition

- Used to tag images in software, security
- Preprocessing

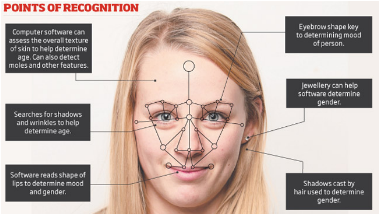


Facial recognition

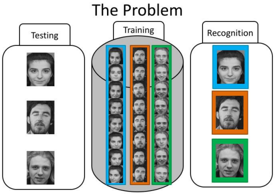
- Used to tag images in software, security
- Preprocessing



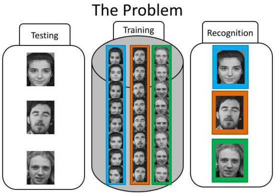
- Feature extraction

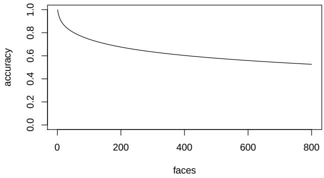


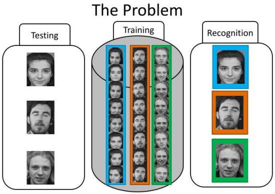
Accuracy vs. number of classes

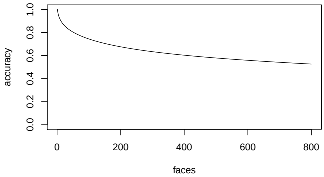


Accuracy vs. number of classes







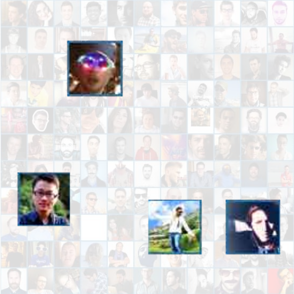


How does the accuracy scale with the number of classes (faces)?

Setup

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

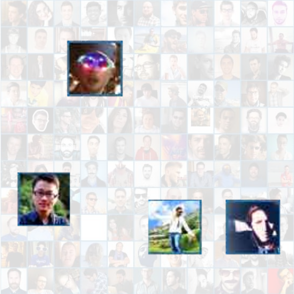
2. Subsample k labels, $y_1, \ldots, y_k$



Setup

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

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

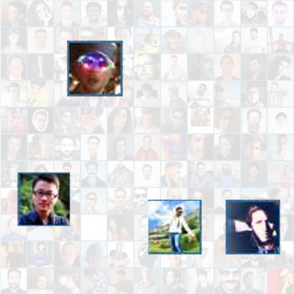


3. Collect training and test data $x_i^{(j)}$ (faces) for labels (people) $\{y_1, \ldots, y_k\}$.

Setup

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

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

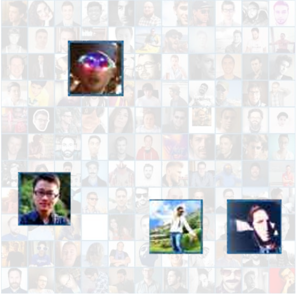


3. Collect training and test data $x_i^{(j)}$ (faces) for labels (people) $\{y_1, \ldots, y_k\}$.
4. Train a classifier and compute test error.

Setup

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

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



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

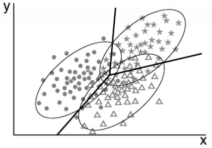
4. Train a classifier and compute test error.

Can we analyze how error depends on k?

Key assumption: marginal classifier

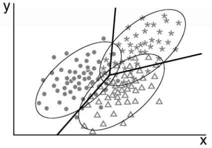
- The classifier is *marginal* if it learns a model *independently* for each class.

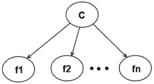
Key assumption: marginal classifier

- The classifier is *marginal* if it learns a model *independently* for each class.
- Examples: LDA/QDA 

,

Key assumption: marginal classifier

- The classifier is *marginal* if it learns a model *independently* for each class.
- Examples: LDA/QDA 

, naïve Bayes 

- Non-marginal classifiers: Multinomial logistic, multilayer neural networks, k-nearest neighbors

Definitions

$\hat{F}_{y^{(i)}}$ is the empirical distribution obtained from the training data for label $y^{(i)}$.

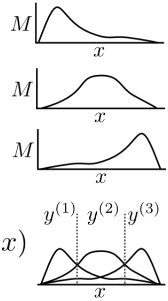
Classification Rule

$$M_{y^{(1)}}(x) = \mathcal{M}(\hat{F}_{y^{(1)}})(x)$$

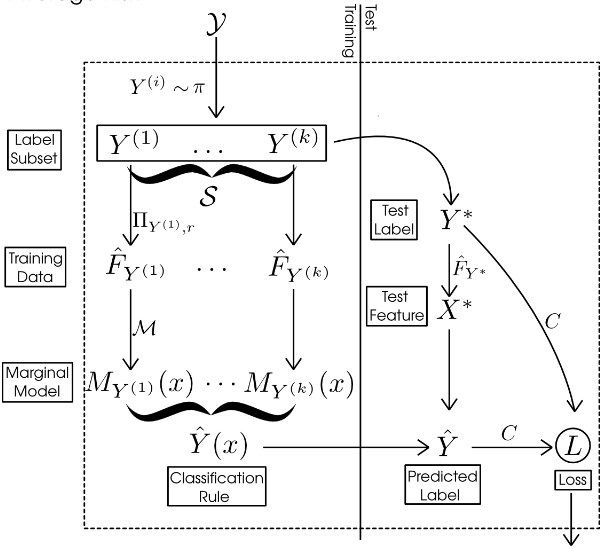
$$M_{y^{(2)}}(x) = \mathcal{M}(\hat{F}_{y^{(2)}})(x)$$

$$M_{y^{(3)}}(x) = \mathcal{M}(\hat{F}_{y^{(3)}})(x)$$

$$\hat{Y}(x) = \text{argmax}_{y \in \mathcal{S}} M_y(x)$$



Average Risk



$$\text{Risk} = \mathbb{E}[L]$$

Theoretical Result

Theorem. (Z., Achanta, Benjamini.) Suppose π, $\{F_y\}_{y \in \mathcal{Y}}$ and marginal classifier $\mathcal{F}$ satisfy *(some regularity condition)*. Then, there exists some function $\bar{D}(u)$ on $[0,1] \to [0,1]$ such that the k-class average risk is given by

$$\text{AvRisk}_k = (k-1) \int \bar{D}(u) u^{k-2} du. \quad (1)$$

Theoretical Result

Theorem. (Z., Achanta, Benjamini.) Suppose π, $\{F_y\}_{y \in \mathcal{Y}}$ and marginal classifier $\mathcal{F}$ satisfy *(some regularity condition)*. Then, there exists some function $\bar{D}(u)$ on $[0,1] \to [0,1]$ such that the k-class average risk is given by

$$\text{AvRisk}_k = (k-1) \int \bar{D}(u) u^{k-2} du. \tag{1}$$

What is this $\bar{D}(u)$ function? We will explain in the following toy example...

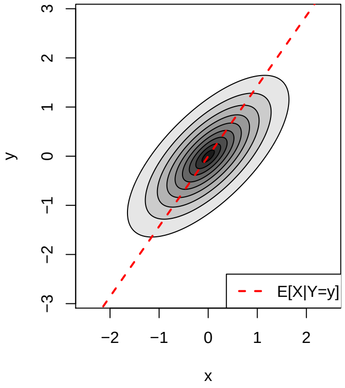
Toy example

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

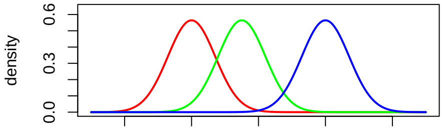
Toy example

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

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

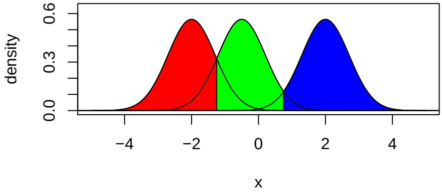


Toy example



- 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.

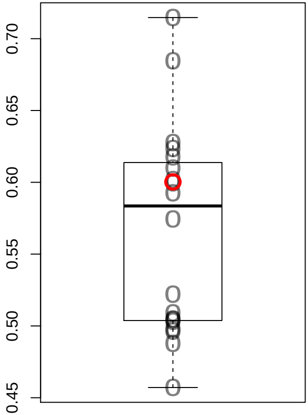
Toy example

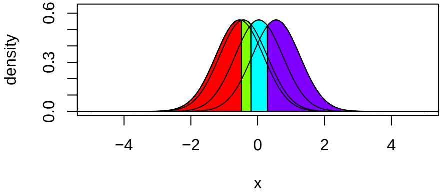


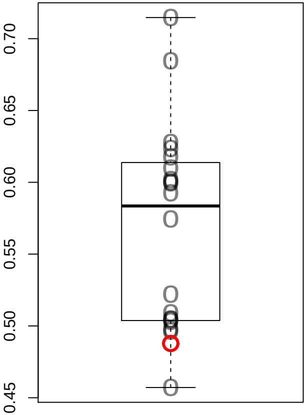
- The *Bayes Risk* is the expected test error of the Bayes rule,

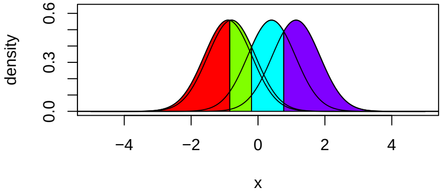
$$\frac{1}{k}\sum_{i=1}^{k} \Pr[\hat{y}(x) \neq Y | Y = y_i]$$

Toy example

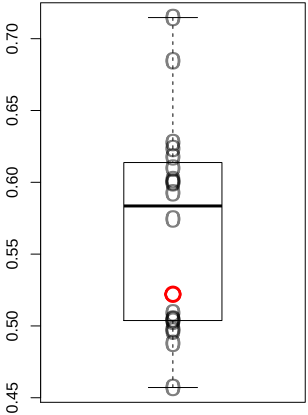


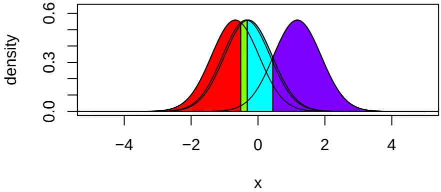



0.70
0.65
0.60
0.55
0.50
0.45

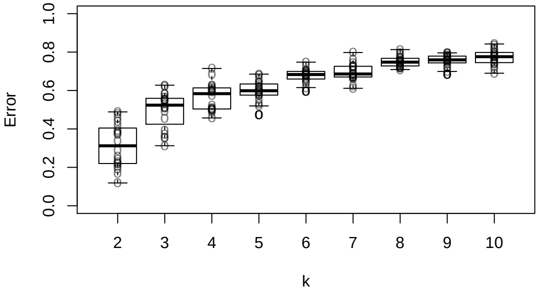

density
0.6
0.3
0.0
−4
−2
0
2
4
x

Toy example

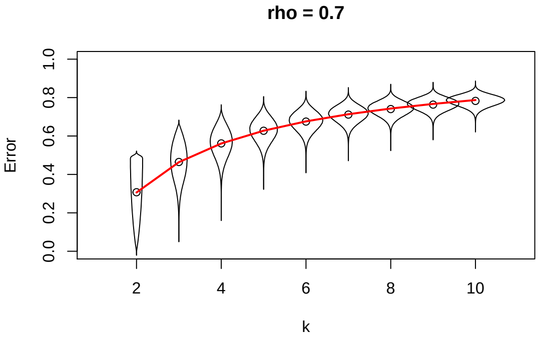




Toy example



Toy example

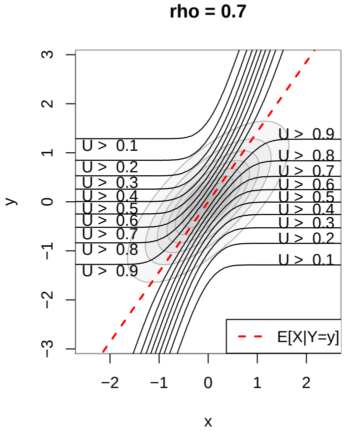


Defining the U-function

Define $U_x(y)$ as follows:

- Suppose we have test instance (face) x whose true label (person) is y.
- Let Y' be a random *incorrect* label (person).
- Use the classifier to guess whether x belongs to y or Y'.
- Define $U_x(y)$ as the probabilility of success (randomizing over training data).

Toy example



$$U_y(x) = \Pr[d(x, \rho Y') > d(x, \rho y)], \text{ for } Y' \sim N(0, 1).$$

Defining the $\bar{D}(u)$

- Define random variable as $U_Y(X)$ for (Y, X) drawn from the joint distribution.

Defining the $\bar{D}(u)$

- Define random variable as $U_Y(X)$ for (Y, X) drawn from the joint distribution.
- $\bar{D}(u)$ is the cumulative distribution function of U,

$$\bar{D}(u) = \Pr[U_Y(X) \leq u].$$

Defining the $\bar{D}(u)$

- Define random variable as $U_Y(X)$ for (Y, X) drawn from the joint distribution.
- $\bar{D}(u)$ is the cumulative distribution function of U,

$$\bar{D}(u) = \Pr[U_Y(X) \leq u].$$