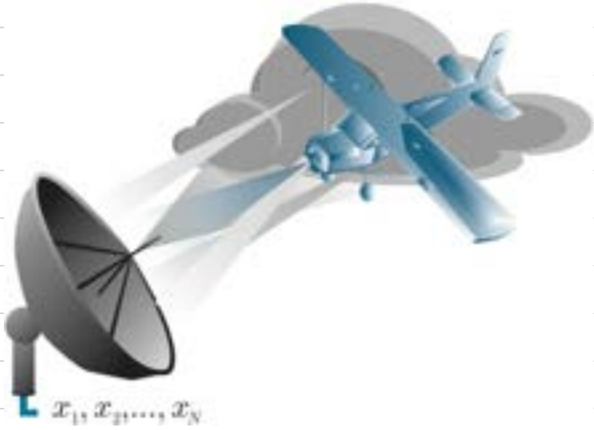
Data-Driven Binary Hypothesis Testing

Example 1



SINGLE DATASET $x_1, x_2, \ldots, x_N$

TWO SCENARIOS (Hypotheses)

$$\mathrm{H}_0 : \quad x_n \sim \text{pure noise}$$

$$\mathrm{H}_1 : \quad x_n \sim \text{noise} + \underbrace{\text{reflection}}_{\text{Presence of airplane}}$$

Using the measured data decide which hypothesis is the most likely to have generated the measurements.

Example 2

Interested in distinguishing between **handwritten numerals** “4” and “9”

Data is an image

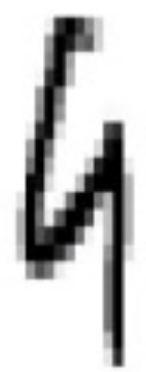
Single image ⇒ Two scenarios

Distinguish (classify) between “4” and “9”

Challenging Problem?

From MNIST

labeled as “9”



labeled as “4”

Hypothesis Testing – Decision Making – Classification

SAME MATHEMATICAL PROBLEM

CAN WE FIND OPTIMUM SOLUTION???

Mathematical Formulation

Need to find a proper way to formulate our problem

Denote $X = \{x_1, \ldots, x_N\}$ the measured data. We assume that X is a realization of a random vector $\mathcal{X}$.

Random vectors, exactly like random variables, are described by probability densities

To be able to distinguish the hypotheses $\mathcal{X}$ must have a different random behavior per hypothesis

$$\mathsf{H}_0: \quad \mathcal{X} \sim \mathsf{f}_0(X), \ \mathbb{P}(\mathsf{H}_0)$$

$$\mathsf{H}_1: \quad \mathcal{X} \sim \mathsf{f}_1(X), \ \mathbb{P}(\mathsf{H}_1)$$

$\mathbb{P}(\mathsf{H}_0), \mathbb{P}(\mathsf{H}_1)$ is our prior knowledge regarding frequency of occurrence of each hypothesis

The Optimum Bayes Test

Every decision mechanism equivalent to a Decision Function $D(X) \in \{0,1\}$

$$D(X) = \begin{cases} 0 & \text{when for } X \text{ we decide } \mathsf{H}_0 \\ 1 & \text{when for } X \text{ we decide } \mathsf{H}_1 \end{cases}$$

Can we optimize $D(X)$?

In what sense ????

We do not like making errors in our decisions!!!!
⇒ MINIMIZE THE ERROR PROBABILITY

$$\mathbb{P}_{\mathrm{E}} = \mathbb{P}\big(D(X) = 1 | \mathsf{H}_0\big)\mathbb{P}(\mathsf{H}_0) + \mathbb{P}\big(D(X) = 0 | \mathsf{H}_1\big)\mathbb{P}(\mathsf{H}_1)$$

With very simple Math we can show that the optimum decision function has the following form

$$D_o(X) = \begin{cases} 1 & \text{when } \dfrac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} > \dfrac{\mathbb{P}(\mathsf{H}_0)}{\mathbb{P}(\mathsf{H}_1)} \\ 0 & \text{when } \dfrac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} < \dfrac{\mathbb{P}(\mathsf{H}_0)}{\mathbb{P}(\mathsf{H}_1)} \end{cases}$$

Optimum decision needs ONLY the Likelihood Ratio Function

$$\mathsf{L}(X) = \frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}$$

and can be written as

$$\mathsf{L}(X) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \frac{\mathbb{P}(\mathsf{H}_0)}{\mathbb{P}(\mathsf{H}_1)} \iff \mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)} \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1$$

We can replace the **vector** X with the **scalar** $\mathsf{L}(X)$ without loosing anything from optimality.

$\mathsf{L}(X)$ is a Sufficient Statistic for the Hypothesis testing problem.

If $\omega(r),\ r \geq 0$ is strictly increasing then

$$\omega\left(\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)}\right) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1)$$

IS ALSO OPTIMUM!!!

Common $\omega(r)$ functions:

$$\omega(r) = \log r \;\Rightarrow\; \text{log-likelihood ratio function}$$

$$\omega(r) = \frac{r}{r+1} \;\Rightarrow\; \text{posterior probability function}$$

Multiple Hypotheses

We can easily extend to more than two hypotheses

$$\mathsf{H}_0 : \quad \mathcal{X} \sim \mathsf{f}_0(X), \mathbb{P}(\mathsf{H}_0)$$
$$\mathsf{H}_1 : \quad \mathcal{X} \sim \mathsf{f}_1(X), \mathbb{P}(\mathsf{H}_1)$$
$$\vdots$$
$$\mathsf{H}_{K-1} : \quad \mathcal{X} \sim \mathsf{f}_{K-1}(X), \mathbb{P}(\mathsf{H}_{K-1})$$

Decision function $D(X) \in \{0, 1, \ldots, K-1\}$

Optimum Decision function:

$$D_o(X) = \arg\max_i \{\mathsf{f}_i(X)\mathbb{P}(\mathsf{H}_i)\}$$

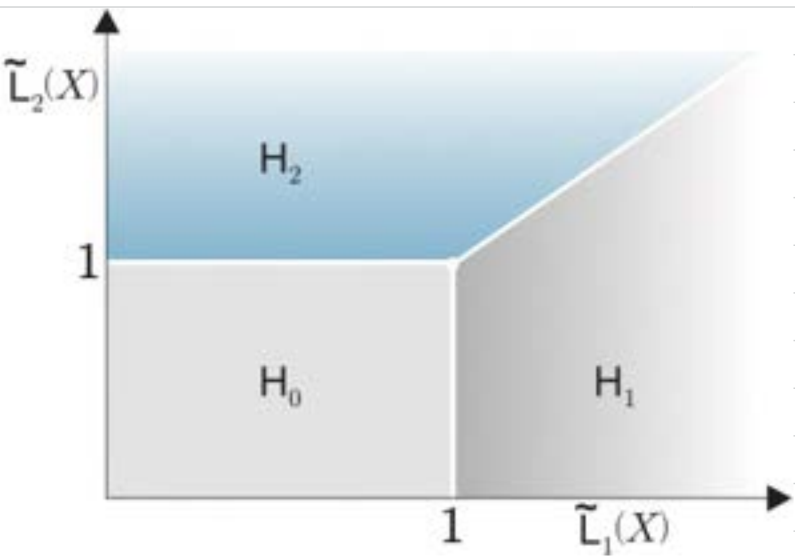
$$H_0 : \quad \mathcal{X} \sim f_0(X), \mathbb{P}(H_0)$$

$$H_1 : \quad \mathcal{X} \sim f_1(X), \mathbb{P}(H_1)$$

$$H_2 : \quad \mathcal{X} \sim f_2(X), \mathbb{P}(H_2)$$

$\mathbb{P}(H_0) + \mathbb{P}(H_1) + \mathbb{P}(H_2) = 1$

$$\tilde{L}_1(X) = \frac{f_1(X)\mathbb{P}(H_1)}{f_0(X)\mathbb{P}(H_0)}, \qquad \tilde{L}_2(X) = \frac{f_2(X)\mathbb{P}(H_2)}{f_0(X)\mathbb{P}(H_0)}$$



What if densities are UNKNOWN????

Can we come up with DATA-DRIVEN version of the optimum test???

Basic Tools

Neural Netwoks

A class of special parametric functions

$$u(X, \theta), \quad \theta: \text{ network parameters}$$

FACT: If $v(X)$ any function then we can approximate it ARBITRARILY CLOSE by a neural network of sufficiently high order

Searching over θ to define a neural network $u(X, \theta)$, when the size of the network tends to infinity

IS EQUIVALENT TO SEARCH OVER A GENERAL FUNCTION $v(X)$

Law of Large Numbers (LLN)

$\mathcal{X}$ random and $\{X_1, X_2, \ldots, X_N\}$ realizations

Let $G(X)$ be a deterministic function, then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} G(X_i) = \mathbb{E}_{\mathcal{X}}[G(\mathcal{X})]$$

$$= \int G(X)\mathsf{f}(X)dX$$

Gradient Descent

Deterministic function $J(\theta)$. Interested in

$$\min_{\theta} J(\theta)$$

We can use

$$\theta_t = \theta_{t-1} - \mu \nabla_\theta J(\theta_{t-1}), \quad \mu > 0$$

Stochastic Gradient Descent

$$J(\theta) = \mathbb{E}_{\mathcal{X}}[G(\mathcal{X}, \theta)]$$

Instead of $\mathsf{f}(X)$ we have $\{X_1, \ldots, X_N\}$ then

$$\theta_t = \theta_{t-1} - \mu \nabla_\theta G(X_t, \theta_{t-1}), \quad \mu > 0$$

Problem of Interest

Two different datasets (e.g. cats/dogs, "4"/"9")

$\mathsf{H}_0: \quad X_1^0, X_2^0, \ldots, X_{N_0}^0 \quad$ (dogs or "4"s)

$\mathsf{H}_1: \quad X_1^1, X_2^1, \ldots, X_{N_1}^1 \quad$ (cats or "9"s)

Assumptions

There exist probability densities $\mathsf{f}_0(X), \mathsf{f}_1(X)$ for $\mathsf{H}_0, \mathsf{H}_1$ that are considered unknown and where dataset $\{X_1^i, \ldots, X_{N_i}^i\}$ is sampled from $\mathsf{f}_i(X)$

There exist prior probabilities $\mathbb{P}(\mathsf{H}_0), \mathbb{P}(\mathsf{H}_1)$ for $\mathsf{H}_0, \mathsf{H}_1$ that are considered unknown with the number of samples being consistent with the priors in the sense

$$\frac{N_i}{N_0 + N_1} \approx \mathbb{P}(\mathsf{H}_i)$$

For every new realization X I would like to decide whether it is from H_0 or H_1

Classical Solution

Design a function which takes the value -1 when X from H_0 and the value 1 when X from H_1

Let the function we are looking for be represented as a neural network $u(X,\theta)$. Then we find the optimum θ by solving the following optimization

$$\min_{\theta}\left\{\sum_{i=1}^{N_0}\big(-1-u(X_i^0,\theta)\big)^2+\sum_{j=1}^{N_1}\big(1-u(X_j^1,\theta)\big)^2\right\}$$

Gradient Descent $\Rightarrow \theta_o \Rightarrow u(X,\theta_o)$

How do I classify?

For every new realization X we observe $u(X,\theta_o)\neq\pm1$. We therefore use

$$u(X,\theta_o) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 0$$

Is this a "good" decision strategy?

Does it approximate the optimum test?

If we have an infinite number of data do we recover the optimum? (CONSISTENCY)

If a strategy is not consistent then for sufficiently large data size an alternative consistent strategy will outperform it!

Asymptotic Analysis

We let $N_0, N_1 \to \infty$. Also we let the size of the neural network $u(X, \theta)$ tend to ∞. The latter suggests that $u(X, \theta)$ can become any function $v(X)$.

$$\min_{\theta}\left\{\sum_{i=1}^{N_0}\big(-1-u(X_i^0,\theta)\big)^2+\sum_{j=1}^{N_1}\big(1-u(X_j^1,\theta)\big)^2\right\}$$

$$\min_{\theta}\left\{\frac{1}{N_0+N_1}\sum_{i=1}^{N_0}\big(1+u(X_i^0,\theta)\big)^2+\frac{1}{N_0+N_1}\sum_{j=1}^{N_1}\big(1-u(X_j^1,\theta)\big)^2\right\}$$

$$\min_{\theta}\left\{\frac{N_0}{N_0+N_1}\frac{1}{N_0}\sum_{i=1}^{N_0}\big(1+u(X_i^0,\theta)\big)^2+\frac{N_1}{N_0+N_1}\frac{1}{N_1}\sum_{j=1}^{N_1}\big(1-u(X_j^1,\theta)\big)^2\right\}$$

Consider $N_0, N_1 \to \infty$, $u(X,\theta) \to v(X)$,

$\min\limits_{\theta} \to \min\limits_{v(X)}$

Asymptotically, optimization is equivalent

$$\min_{v(X)}\left\{\mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\Big[\big(1+v(X)\big)^2\Big]+\mathbb{P}(\mathsf{H}_1)\mathbb{E}_1\Big[\big(1-v(X)\big)^2\Big]\right\}$$

$$\int \mathbb{P}(\mathsf{H}_0)\big(1+v(X)\big)^2\mathsf{f}_0(X)dX+\int \mathbb{P}(\mathsf{H}_1)\big(1-v(X)\big)^2\mathsf{f}_1(X)dX$$

$$\mathsf{f}_1(X)=\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\mathsf{f}_0(X)=\mathsf{L}(X)\mathsf{f}_0(X)$$

$$\int\left\{\mathbb{P}(\mathsf{H}_0)\big(1+v(X)\big)^2+\mathbb{P}(\mathsf{H}_1)\big(1-v(X)\big)^2\mathsf{L}(X)\right\}\mathsf{f}_0(X)dX$$

$$\min_v \left\{ \mathbb{P}(\mathsf{H}_0)(1+v)^2 + \mathbb{P}(\mathsf{H}_1)(1-v)^2 \mathsf{L} \right\}$$

The optimum solution is

$$v_o(X) = \frac{\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)} - 1}{\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)} + 1} = \omega\left(\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)}\right)$$

where $\omega(r) = \frac{r-1}{r+1}$, strictly increasing

So $$v_o(X) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1) = 0$$

is equivalent to the optimum test!!!!

We do not have $v_o(X)$. Instead we have a neural network $u(X, \theta_o)$, an approximation of $v_o(X)$.

Our test MUST HAVE THE FORM

$$u(X, \theta_o) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 0$$

General Class of Optimization Problems

$$\min_{v(X)} \left\{ \mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\left(1+v(X)\right)^2\right] + \mathbb{P}(\mathsf{H}_1)\mathbb{E}_1\left[\left(1-v(X)\right)^2\right] \right\}$$

Propose the following cost function

$$\mathcal{G}(v) = \mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\phi\left(v(X)\right)\right] + \mathbb{P}(\mathsf{H}_1)\mathbb{E}_1\left[\psi\left(v(X)\right)\right]$$

The two functions $\phi(z), \psi(z)$ depend on scalar z

Select $\phi(z), \psi(z)$, so that

$$\min_{v(X)} \mathcal{G}(v)$$

has as solution $v_o(X) = \omega\left(\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)}\right)$

for a pre-specified strictly increasing

$\omega(r), r \geq 0$

THEOREM

Select your favorite strictly increasing $\omega(r)$.

Select $\psi(z)$ so that $\psi'(z) < 0$.

Define $\quad \phi'(z) = -\omega^{-1}(z)\psi'(z)$, then

$$\arg\min_{v(X)} \left\{ \mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\phi\big(v(X)\big)\right] + \mathbb{P}(\mathsf{H}_1)\mathbb{E}_1\left[\psi\big(v(X)\big)\right] \right\}$$

$$= v_o(X) = \omega\left(\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)}\right)$$

The test

$$v_o(X) = \omega\left(\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_1)}\right) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1)$$

is optimum

Data-Driven Version

$$v(X) \Rightarrow u(X, \theta)$$

$$\min_{v(X)} \Rightarrow \min_{\theta}$$

$$\mathbb{E}[\,] \Rightarrow \frac{1}{N}\sum_{i=1}^{N}$$

$$J(\theta) = \sum_{i=1}^{N_0} \phi\big(u(X_i^0, \theta)\big) + \sum_{j=1}^{N_1} \psi\big(u(X_j^1, \theta)\big)$$

$$\theta_o = \arg\min_{\theta} J(\theta)$$

$$u(X, \theta_o) \text{ approximates } v_o(X) = \omega\left(\mathsf{L}(X)\frac{\mathbb{P}(\mathsf{H}_1)}{\mathbb{P}(\mathsf{H}_0)}\right)$$

The test we apply is

$$u(X, \theta_o) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1)$$

which is CONSISTENT

Examples

$$\omega(r) = r, \qquad \omega(1) = 1$$

$$\psi'(z) = -1$$

$$\psi(z) = -z, \quad \phi(z) = \tfrac{1}{2}z^2, \quad \text{Mean Square}$$

$$\omega(r) = \log r, \qquad \omega(1) = 0$$

$$\psi'(z) = -e^{-0.5z^2}$$

$$\psi(z) = e^{-0.5z^2}, \quad \phi(z) = e^{0.5z^2}, \quad \text{Exponential}$$

$$\omega(r) = \frac{r}{r+1}, \qquad \omega(1) = 0.5$$

$$\psi'(z) = -\frac{1}{z}$$

$$\psi(z) = -\log z, \quad \phi(z) = -\log(1-z), \quad \text{Cross Entropy}$$

Each optimization problem produces a different function $u(X, \theta_o)$ and therefore classifier

REMARK:

Not all consistent classifiers perform the same!!!!

Decide Between "4" and "9" (MNIST)

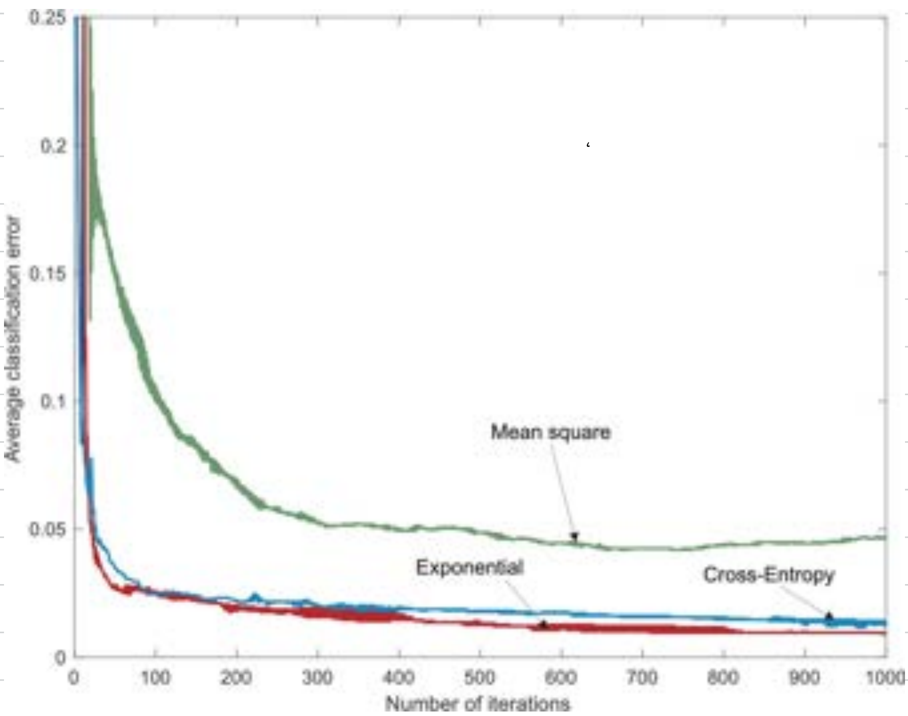
$N_0 = N_1 = 5500$ (Training data)

X is image $28 \times 28 \to 784 \times 1$

$u(X, \theta)$ Full neural network $784 \times 300 \times 1$ with 236,584 parameters (ReLU)

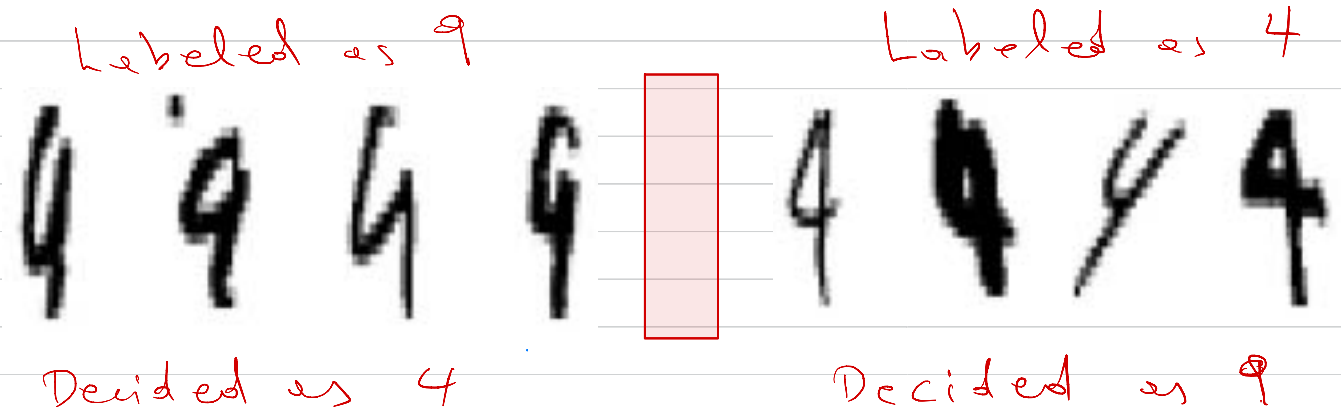
Use gradient descent to compute θ_o

At each iteration we have θ_t and $u(X, \theta_t)$
We apply it to testing data (983 "4" ad 1009 "9")
Observe evolution of error percentage with iterations



Mean Square, significantly worse, because dynamic range of $\mathsf{L}(X)$ is larger than the the range of $\log \mathsf{L}(X)$ or $\frac{\mathsf{L}(X)}{\mathsf{L}(X)+1}$

Examples of decision (classification) error for Exponential Method



MAJOR CHALLENGES

Be able to decide which optimization is appropriate

Extension to the multi-hypothesis case

Relate network size to optimization problem and data size