DATA DRIVEN TECHNIQUES

Lecture 11

Data-Driven Decision Making

Outline

- Hypothesis Testing
  - Mathematical Formulation
  - Data Driven Approach
  - The Consistency Property
- Likelihood Ratio Estimation
  - Optimization Problems with Consistent Solutions
  - Data Driven Implementation
- Detection in Time Series
  - I.i.d. processes
  - Markov processes

Hypothesis Testing

Mathematical Formulation

For a random vector X we assume the following two hypotheses

$$\begin{aligned} \mathsf{H}_0 :&\quad X \sim \mathsf{f}_0(X),\ \ \mathbb{P}(\mathsf{H}_0) \\ \mathsf{H}_1 :&\quad X \sim \mathsf{f}_1(X),\ \ \mathbb{P}(\mathsf{H}_1) \end{aligned}$$

For every X need to decide if it comes from H_0 or H_1

Decide using a *Decision Function* $\mathsf{D}(X) \in \{0, 1\}$

Would like to optimize $\mathsf{D}(X)$

Plethora of applications in diverse scientific fields!!!

Bayesian Approach

Minimize decision error probability

$$\min_{\mathsf{D}} \Big\{ \mathbb{P}(\mathsf{D} = 1 | \mathsf{H}_0)\mathbb{P}(\mathsf{H}_0) + \mathbb{P}(\mathsf{D} = 0 | \mathsf{H}_1)\mathbb{P}(\mathsf{H}_1) \Big\}$$

$$\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \frac{\mathbb{P}(\mathsf{H}_0)}{\mathbb{P}(\mathsf{H}_1)} \quad \equiv \quad \frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)} \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1$$

For $\omega(\mathsf{r})$ strictly increasing

$$\boxed{\mathsf{r}(X) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1 \quad \equiv \quad \omega\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1),} \quad \mathsf{r}(X) = \frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)}$$

Neyman-Pearson Approach

$$\begin{aligned} \mathsf{H}_0 : &\quad X \sim \mathsf{f}_0(X), \ \cancel{\mathbb{P}(\mathsf{H}_0)} \\ \mathsf{H}_1 : &\quad X \sim \mathsf{f}_1(X), \ \cancel{\mathbb{P}(\mathsf{H}_1)} \end{aligned}$$

Maximize detection probability $\mathbb{P}(\mathsf{D} = 1 | \mathsf{H}_1)$

subject to false alarm probability constraint $\mathbb{P}(\mathsf{D} = 1 | \mathsf{H}_0) \leq \alpha$

$$\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \lambda, \quad \mathbb{P}\left(\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} \geq \lambda \middle| \mathsf{H}_0 \right) = \alpha$$

For $\omega(\mathsf{r})$ strictly increasing

$$\boxed{\omega\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \eta, \quad \mathbb{P}\Big(\omega\big(\mathsf{r}(X)\big) \geq \eta \Big| \mathsf{H}_0 \Big) = \alpha,} \qquad \mathsf{r}(X) = \frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}$$

Data Driven Approach

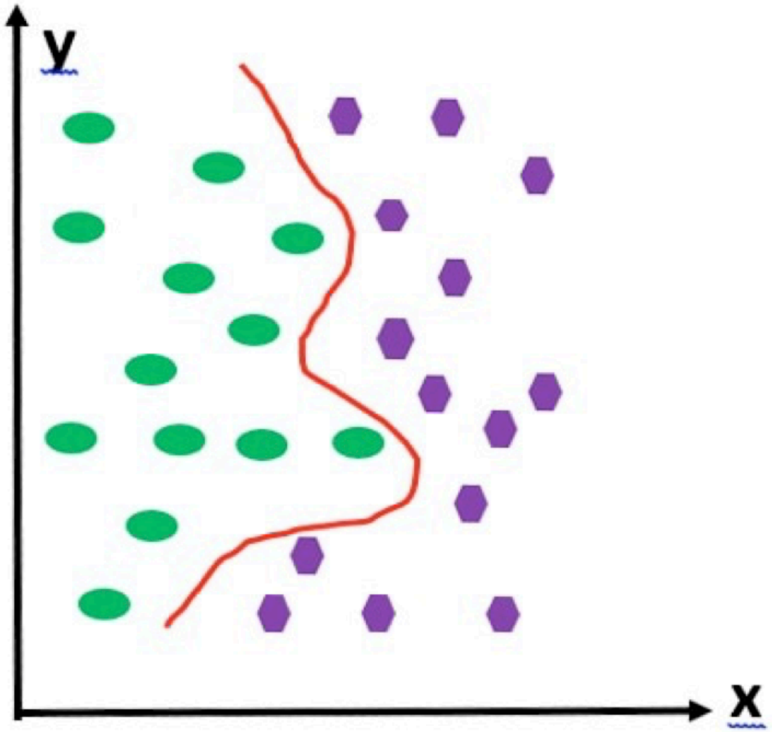
$$\text{H}_0 : \quad X \sim \cancel{\text{f}_0(X)}, \quad \cancel{\mathbb{P}(\text{H}_0)}$$
$$\text{H}_1 : \quad X \sim \cancel{\text{f}_1(X)}, \quad \cancel{\mathbb{P}(\text{H}_1)}$$

$$X_1^0 \;\; X_2^0 \;\; \dots \;\; X_{n_0}^0$$
$$X_1^1 \;\; X_2^1 \;\; \dots \;\; X_{n_1}^1$$

Sampled from f_0

Sampled from f_1

$$\mathbb{P}(\text{H}_i) \approx \frac{n_i}{n_0 + n_1}$$



Design border to separate the two datasets

What is the best border ?

$$\text{All } X : \; \frac{\text{f}_1(X)\mathbb{P}(\text{H}_1)}{\text{f}_0(X)\mathbb{P}(\text{H}_0)} = 1$$

Instead of a “border”, design a decision like function $\mathsf{v}(X)$

$$\mathsf{v}(X) = \begin{cases} -1 & \text{when } X \text{ from } \mathsf{H}_0 \\ 1 & \text{when } X \text{ from } \mathsf{H}_1. \end{cases}$$

Use parametric family of functions $\mathsf{u}(X,\theta)$ and optimize θ solving

$$\mathsf{J}(\theta) = \frac{1}{n_0+n_1}\left\{\sum_{i=1}^{n_0}\Big(-1-\mathsf{u}(X_i^0,\theta)\Big)^2 + \sum_{j=1}^{n_1}\Big(1-\mathsf{u}(X_j^1,\theta)\Big)^2\right\}$$

$$\min_{\theta} \mathsf{J}(\theta) \;\Rightarrow\; \theta_{\mathsf{o}} \;\Rightarrow\; \mathsf{u}(X,\theta_{\mathsf{o}})$$

For every X to test decide as follows: $\mathsf{u}(X,\theta_{\mathsf{o}}) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 0$

Works “well”!! Why??

Understanding using Asymptotic Analysis

$$n_0, n_1 \to \infty, \qquad \mathsf{u}(X,\theta) \to \mathsf{v}(X)$$

$$\boxed{\mathsf{J}(\theta)} = \boxed{\frac{n_0}{n_0+n_1}} \boxed{\frac{1}{n_0}\sum_{i=1}^{n_0}} \Big(1 + \boxed{\mathsf{u}(X_i^0,\theta)}\Big)^2 + \frac{n_1}{n_0+n_1}\,\frac{1}{n_1}\sum_{j=1}^{n_1}\Big(1 - \mathsf{u}(X_j^1,\theta)\Big)^2$$

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

$$\boxed{\min_\theta \mathsf{J}(\theta)} \to \boxed{\min_{\mathsf{v}} \mathsf{J}(\mathsf{v})}$$

$$\theta_o \Rightarrow \mathsf{u}(X,\theta_o) \approx \mathsf{v}_o(X)$$

$$\mathbb{E}_1\left[\Big(1-\mathsf{v}(X)\Big)^2\right] = \mathbb{E}_0\left[\Big(1-\mathsf{v}(X)\Big)^2\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\right]$$

$$\mathsf{J}(\mathsf{v}) = \mathbb{P}(\mathsf{H}_0)\mathbb{E}_0\left[\Big(1+\mathsf{v}(X)\Big)^2 + \mathsf{r}(X)\Big(1-\mathsf{v}(X)\Big)^2\right] \qquad \mathsf{r}(X) = \frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)}$$

minimize for each X

$$\mathsf{v}_\mathsf{o}(X) = \frac{\mathsf{r}(X)-1}{\mathsf{r}(X)+1} = \omega\big(\mathsf{r}(X)\big), \quad \text{where } \omega(\mathsf{r}) = \frac{\mathsf{r}-1}{\mathsf{r}+1} \text{ strictly increasing}$$

Test equivalent to Bayes: $\mathsf{v}_\mathsf{o}(X) = \omega\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1) = 0 \quad \Rightarrow \quad \mathsf{u}(X,\theta_\mathsf{o}) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 0$

Equivalence in the limit

Consistency (with respect to the Bayes test)

$$\min_{\theta} \mathsf{J}(\theta) = \min_{\theta}\left\{ \sum_{i=1}^{n_0}\Big(1+\mathsf{u}(X_i^0,\theta)\Big)^2 + \sum_{j=1}^{n_1}\Big(1-\mathsf{u}(X_j^1,\theta)\Big)^2 \right\}$$

$$\Rightarrow\ \theta_{\mathsf{o}}\ \Rightarrow\ \mathsf{u}(X,\theta_{\mathsf{o}})$$

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

$$\Rightarrow\ \mathsf{v}_{\mathsf{o}}(X)\ =\ \omega\big(\mathsf{r}(X)\big)$$

$$\boxed{\text{Expect: } \mathsf{u}(X,\theta_{\mathsf{o}}) \approx \omega\big(\mathsf{r}(X)\big)}$$

Optimum Test: $\omega\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1)$, Close to Optimum: $\mathsf{u}(X,\theta_{\mathsf{o}}) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega(1)$

Develop data driven methods for estimation of $\omega\big(\mathsf{r}(X)\big)$ for other $\omega(\mathsf{r})$

Consistent tests eventually prevail over inconsistent tests

Likelihood Ratio Estimation

$$\mathsf{r}(X) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1 \;\equiv\; \omega_1\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega_1(1) \;\equiv\; \omega_2\big(\mathsf{r}(X)\big) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega_2(1)$$

$$\boxed{\mathsf{u}(X,\theta_{\mathsf{o}}) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} 1 \;\not\equiv\; \mathsf{u}_1(X,\theta_1) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega_1(1) \;\not\equiv\; \mathsf{u}_2(X,\theta_2) \underset{\mathsf{H}_0}{\overset{\mathsf{H}_1}{\gtreqless}} \omega_2(1)}$$

For function $\omega(\mathsf{r})$ can we define cost

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

so that $\min_{\mathsf{v}} \mathsf{J}(\mathsf{v}) \;\Rightarrow\; \mathsf{v}_{\mathsf{o}}(X) = \omega\big(\mathsf{r}(X)\big)$?

Consider $\omega(r)$ strictly increasing. Denote with I_ω the range of $\omega(r)$.

THEOREM

For $z \in I_\omega$ select $\rho(z)$ strictly negative and define two function $\phi(z), \psi(z)$ for $z \in I_\omega$ as follows:

$$\boxed{\phi'(z) = -\omega^{-1}(z)\rho(z), \qquad \psi'(z) = \rho(z)}$$

then

$$\min_{V} \left\{ \mathbb{P}(H_0)\mathbb{E}_0[\phi(V(X))] + \mathbb{P}(H_1)\mathbb{E}_1[\psi(V(X))] \right\}$$

accepts a unique minimizer $V_0(X)$ that satisfies

$$\boxed{V_0(X) = \omega\big(r(X)\big)}$$

where we recall: $r(X) = \dfrac{\mathbb{P}(H_1) f_1(x)}{\mathbb{P}(H_0) f_0(x)}$ Proof →

$$J(v) = \mathbb{E}_0\left[\mathbb{P}(H_0)\,\phi(v(x))\right] + \mathbb{E}_0\left[\mathbb{P}(H_1)\,\psi(v(x))\,L(x)\right] \qquad L(x) = \frac{f_1(x)}{f_0(x)}$$

change of measure

$$= \mathbb{P}(H_0)\,\mathbb{E}_0\left[\left\{\phi(v(x)) + \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} L(x)\,\psi(v(x))\right\}\right] = \mathbb{P}(H_0)\,\mathbb{E}_0\left[\underbrace{\phi(v(x)) + r(x)\,\psi(v(x))}_{\text{Minimize over } V(x) \text{ for each fixed } X}\right]$$

$$r(x) = \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} L(x)$$

For fixed X $\quad v(x) \sim v, \quad r(x) \sim r \geq 0$ and interested in

$$\min_v \{\phi(v) + r\,\psi(v)\}$$

Take derivative with respect to v and use definition of ϕ', ψ'

$$\phi'(v) + r\,\psi'(v) = 0 \Longrightarrow -\bar{\omega}(v)\rho(v) + r\rho(v) = 0$$

$$\rightarrow \rho(v)\{r - \bar{\omega}(v)\} = 0 \Longrightarrow \text{Unique solution}$$

$$\bar{\omega}^{-1}(V_0) = r \Longrightarrow V_0 = \omega(r)$$

It is also a minimum because derivative is negative for $V < V_0$ and positive for $V > V_0$

EXAMPLES OF FUNCTIONS

A: $w(r) = r \geq 0$ and $I_w = \mathbb{R}_+$ — Estimate likelihood ratio

A1: $\rho(t) = -t^{\alpha}$. then $\phi(t) = \frac{t^{2+\alpha}}{2+\alpha}$, $\psi(t) = -\frac{t^{1+\alpha}}{1+\alpha}$ $\quad \alpha \neq -1, -2$

special case: $\alpha = 0$, $\rho(t) = -1 \longrightarrow \phi(t) = \frac{t^2}{2}$, $\psi(t) = -t$ MSE

A2: $\rho(t) = -\frac{1}{t(1+t)}$ then $\phi(t) = \log(1+t)$, $\psi(t) = \log(1+t^{-1})$

B: $w(r) = \log r$ and $I_w = \mathbb{R}$ — Estimate log-likelihood ratio

B1: $\rho(t) = -e^{-\alpha t}$ then $\phi(t) = \frac{e^{(1-\alpha)t} - 1}{1-\alpha}$, $\psi(t) = \frac{e^{-\alpha t} - 1}{\alpha}$

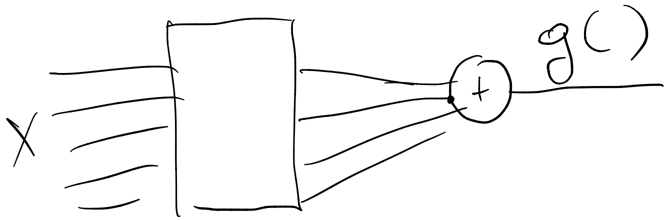
B2: $\rho(t) = -\frac{1}{1+e^{t}}$ then $\phi(t) = \log(1+e^{t})$, $\psi(t) = \log(1+e^{-t})$

C: $w(r) = \frac{r}{r+1}$ and $I_w = (0,1)$ — Estimate Posterior probability

C1: $\rho(t) = -\frac{1}{t}$ $(t \in I_w)$ then $\phi(t) = -\log(1-t)$, $\psi(t) = -\log t$

ATTENTION

We have to make sure that the output of the network $u(x, \theta_0)$ is consistent with the range I_ω. This means that at its output we must add a proper nonlinearity $g(\,)$.



A) $\omega(r) = r$ with $I_\omega = (0, \infty)$ then $g(x) = \max(x, 0)$ ReLU

B) $\omega(r) = \log r$ with $I_\omega \in \mathbb{R}$ then $g(x) = x$ (no nonlinearity)

C $\omega(r) = \dfrac{r}{r+1}$ with $I_\omega \in (0, 1)$ then $g(x) = \dfrac{1}{1+e^{-x}}$ sigmoid

There will be a newer version of the theorem where we are going to incorporate the actual output nonlinearity directly in our analysis!

Examples of functions

A: $\omega(\mathsf{r}) = \mathsf{r} \in \mathbb{R}_+$ (likelihood ratio)

$$\rho(z) = -1, z \geq 0 \;\; \Rightarrow \;\; \phi(z) = \frac{z^2}{2}, \;\; \psi(z) = -z$$

Mean Square

B: $\omega(\mathsf{r}) = \log(\mathsf{r}) \in \mathbb{R}$ (log-likelihood ratio)

$$\rho(z) = -e^{-0.5z} \;\; \Rightarrow \;\; \phi(z) = 2e^{0.5z}, \;\; \psi(z) = 2e^{-0.5z}$$

Exponential

C: $\omega(\mathsf{r}) = \frac{\mathsf{r}}{\mathsf{r}+1} \in [0,1]$ (posterior probability)

$$\rho(z) = -\frac{1}{z}, z \in [0,1] \;\; \Rightarrow \;\; \phi(z) = -\log(1-z), \;\; \psi(z) = -\log(z)$$

Cross Entropy

3

④

Data Driven Implementation

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

$$\mathsf{J}(\theta) = \frac{1}{n_0+n_1}\left\{\sum_{i=1}^{n_0}\phi\big(\mathsf{u}(X_i^0,\theta)\big) + \sum_{j=1}^{n_1}\psi\big(\mathsf{u}(X_j^1,\theta)\big)\right\}$$

$$\mathsf{u}(X,\theta_o) \approx \omega\left(\frac{\mathsf{f}_1(X)\mathbb{P}(\mathsf{H}_1)}{\mathsf{f}_0(X)\mathbb{P}(\mathsf{H}_0)}\right)$$

If instead we define

$$\mathsf{J}(\theta) = \frac{1}{n_0}\sum_{i=1}^{n_0}\phi\big(\mathsf{u}(X_i^0,\theta)\big) + \frac{1}{n_1}\sum_{j=1}^{n_1}\psi\big(\mathsf{u}(X_j^1,\theta)\big)$$

then

$$\mathsf{u}(X,\theta_o) \approx \omega\left(\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\right)$$

NUMERICAL IMPLEMENTATION

Gradient Descent.

$$\theta_t = \theta_{t-1} - h\left\{ \sum_{i=1}^{m_0} \nabla_\theta \phi\left(u(x_i^0, \theta_{t-1})\right) + \sum_{j=1}^{m_1} \nabla_\theta \phi\left(u(x_j^1, \theta_{t-1})\right) \right\}$$

$$= \theta_{t-1} - h\left\{ -\sum_{i=1}^{m_0} \omega^{-1}\left(u(x_i^0, \theta_{t-1})\right) \rho\left(u(x_i^0, \theta_{t-1})\right) \nabla_\theta u(x_i^0, \theta_{t-1}) + \sum_{j=1}^{n_1} \rho\left(u(x_j^1, \theta_{t-1})\right) \nabla_\theta u(x_j^1, \theta_{t-1}) \right\}$$

then $$u(x, \theta_\infty) \simeq \omega\left(\frac{\mathbb{P}(H_1)\, f_1(x)}{\mathbb{P}(H_0)\, f_0(x)}\right)$$

If we normalize each sum with the number of its samples

then $$u(x, \theta_\infty) \simeq \omega\left(\frac{f_1(x)}{f_0(x)}\right)$$

STOCHASTIC GRADIENT DESCENT

We mix randomly the two datasets retaining their labels (0 or 1)

then at iteration t we use X_t

$$\theta_t = \theta_{t-1} + \begin{cases} -\omega^{-1}(u(X_t,\theta_{t-1}))\,\rho(u(X_t,\theta_{t-1}))\,\nabla_\theta u(X_t,\theta_{t-1}) & \text{if } X_t = X_t^0 \\ \rho(u(X_t,\theta_{t-1}))\,\nabla_\theta u(X_t,\theta_{t-1}) & \text{if } X_t = X_t^1 \end{cases}$$

then $$u(X,\theta_\infty) \simeq \omega\left(\frac{\mathbb{P}(H_1)\,h(x)}{\mathbb{P}(H_0)\,f_0(x)}\right)$$

If we multiply the upper term with $\frac{1}{n_0}$ and the lower with $\frac{1}{n_1}$ then

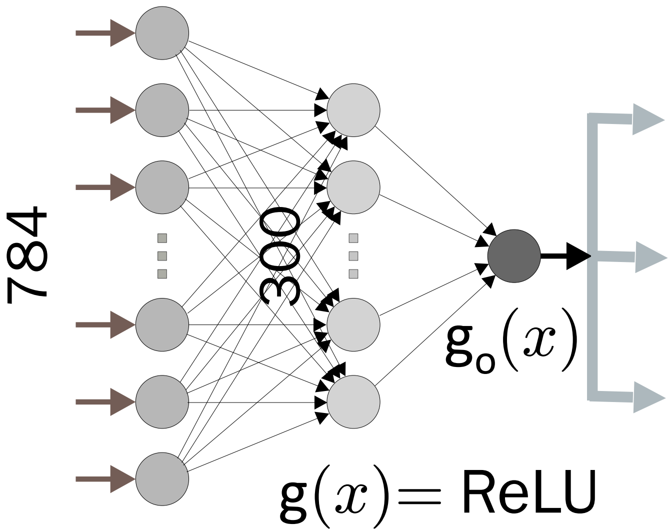
$$\omega(X,\theta_\infty) \simeq \omega\left(\frac{h(x)}{f_0(x)}\right)$$

Example: Classification Problem

From dataset MNIST isolate handwritten numerals 4 and 9

Gray scale images 28 X 28 = 784 pixels. Design classifier using training data. Examine performance using testing data.

Neural network 784 X 300 X 1



Mean square $g_o(x) =$ ReLU, $u_1(X,\theta_1)$

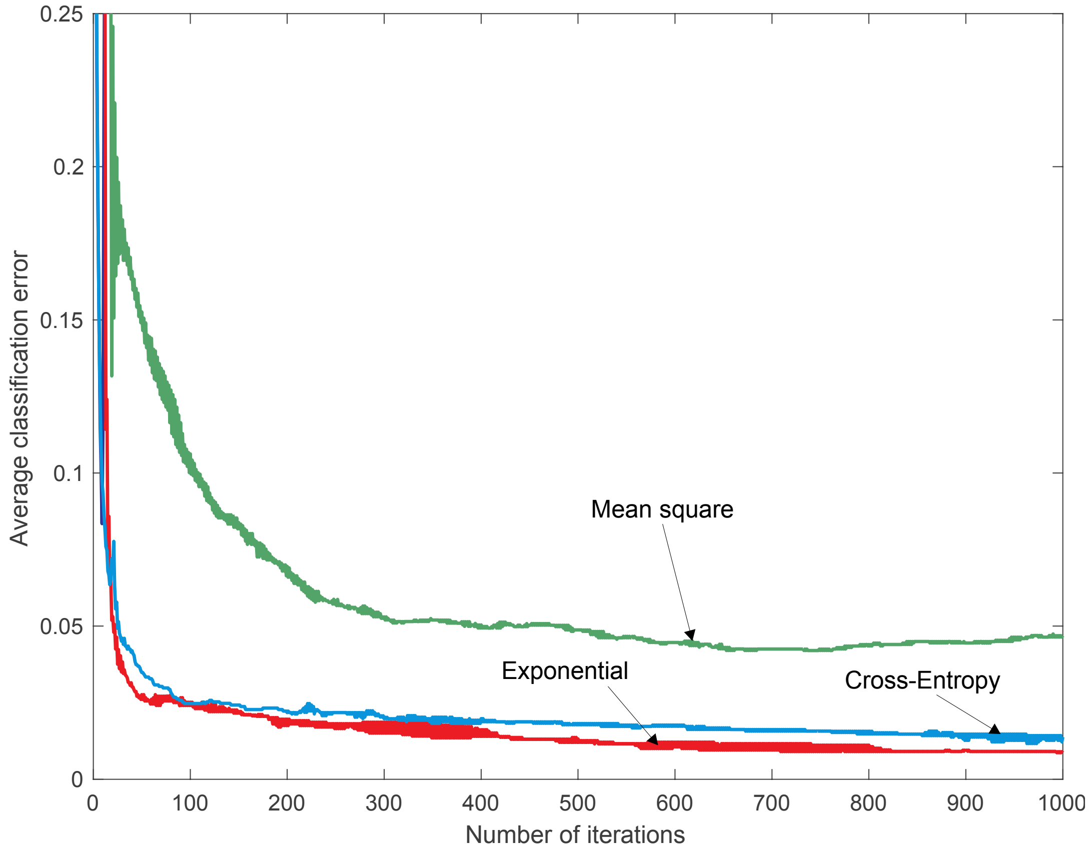
Exponential $g_o(x) = x$, $u_2(X,\theta_2)$

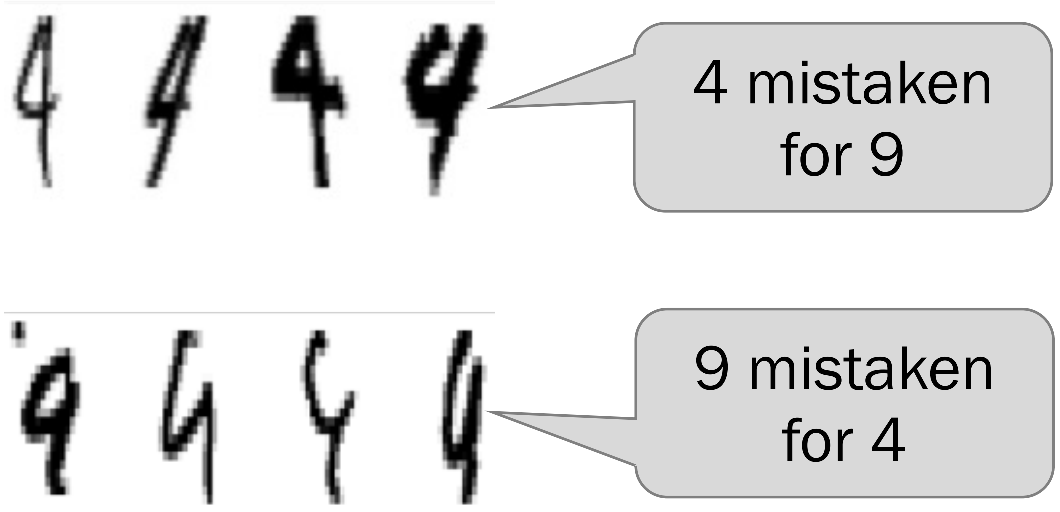
Cross entropy $g_o(x) =$ sigmoid, $u_3(X,\theta_3)$

# of parameters
235801

Training set: 5500 “4” and 5500 “9”. Testing set: 982 “4” and 1009 “9”

$$u_1(X,\theta_1) \underset{H_0}{\overset{H_1}{\gtreqless}} 1, \quad u_2(X,\theta_2) \underset{H_0}{\overset{H_1}{\gtreqless}} 0, \quad u_3(X,\theta_3) \underset{H_0}{\overset{H_1}{\gtreqless}} \frac{1}{2}$$





9

Detection in Time Series

More practically interesting case: Testing of time series { X_1 , X_2 ,...., X_n }

The whole set of measurements under H_0 or H_1

For testing we need likelihood ratio

$$\frac{f_1(X_n,\ldots,X_1)}{f_0(X_n,\ldots,X_1)} = \frac{f_1(X_n|X_{n-1},\ldots,X_1)}{f_0(X_n|X_{n-1},\ldots,X_1)}\frac{f_1(X_{n-1}|X_{n-2},\ldots,X_1)}{f_0(X_{n-1}|x_{X-2},\ldots,X_1)}\cdots\frac{f_1(X_1)}{f_0(X_1)}$$

When i.i.d. under each hypothesis

$$\frac{f_1(X_n,\ldots,X_1)}{f_0(X_n,\ldots,X_1)} = \frac{f_1(X_n)}{f_0(X_n)}\cdots\frac{f_1(X_1)}{f_0(X_1)}$$

Test to be used

$$\boxed{\sum_{i=1}^{n}\log\left(\frac{f_1(X_i)}{f_0(X_i)}\right) \underset{H_0}{\overset{H_1}{\gtreqless}} \eta}$$

13

Interested in estimating $\omega\big(\mathsf{r}(X)\big) = \omega\left(\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\right)$

We are given training data: $\{X_1^0, \ldots, X_{n_0}^0\}$ following H_0

$\{X_1^1, \ldots, X_{n_1}^1\}$ following H_1

For each $\omega(\mathsf{r})$ of interest, minimize corresponding $\mathsf{J}(\theta)$

$$\mathsf{J}(\theta) = \frac{1}{n_0}\sum_{i=1}^{n_0} \phi\big(\mathsf{u}(X_i^0, \theta)\big) + \frac{1}{n_1}\sum_{j=1}^{n_1} \psi\big(\mathsf{u}(X_j^1, \theta)\big)$$

$$\mathsf{u}(X, \theta_{\mathsf{o}}) \approx \omega\left(\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\right)$$

11

$$\text{For } \omega(\mathsf{r}) = \mathsf{r}, \quad \mathsf{u}_1(X, \theta_1) \approx \frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}, \quad \text{use} \quad \log\big(\mathsf{u}_1(X, \theta_1)\big) \quad \text{(Mean Square)}$$

$$\text{For } \omega(\mathsf{r}) = \log(\mathsf{r}), \quad \mathsf{u}_2(X, \theta_2) \approx \log\left(\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}\right), \quad \text{use} \quad \mathsf{u}_2(X, \theta_2) \quad \text{(Exponential)}$$

$$\text{For } \omega(\mathsf{r}) = \frac{\mathsf{r}}{\mathsf{r}+1}, \quad \mathsf{u}_3(X, \theta_3) \approx \frac{\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)}}{\frac{\mathsf{f}_1(X)}{\mathsf{f}_0(X)} + 1}, \quad \text{use} \quad \log\left(\frac{\mathsf{u}_3(X, \theta_3)}{1 - \mathsf{u}_3(X, \theta_3)}\right)$$

(Cross Entropy)

Example: Testing i.i.d. sequences

X_i length 10, $f_0(X) \sim \mathcal{N}(0, I)$

$f_1(X) \sim \mathcal{N}(\frac{1}{\sqrt{10}}[1 \cdots 1]^\intercal, 1.2I)$

Neural Network 10×20×1

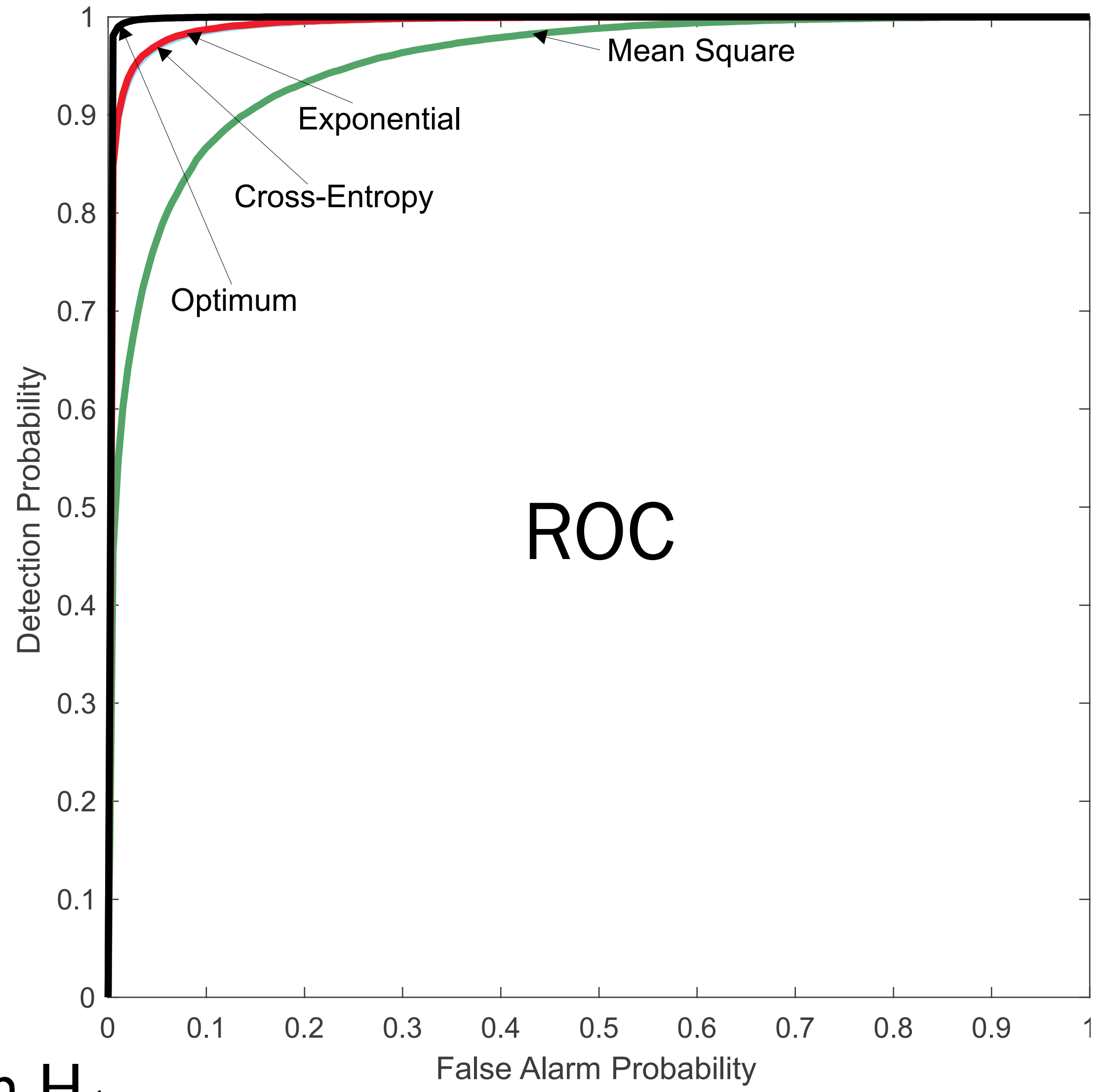
Training data $n_0 = n_1 =$100

Produce $u_1(X, \theta_1), u_2(X, \theta_2), u_3(X, \theta_3)$

Testing for $n =$20 samples

$$\left.\begin{array}{r} \sum_{i=1}^{20} \log\big(u_1(X_i, \theta_1)\big) \\ \sum_{i=1}^{20} u_2(X_i, \theta_2) \\ \sum_{i=1}^{20} \log\left(\frac{u_3(X_i, \theta_3)}{1 - u_3(X_i, \theta_3)}\right) \end{array}\right\} \underset{H_0}{\overset{H_1}{\gtreqless}} \eta$$

100000×20 realizations from H_0 and from H_1



Markovian processes

Consider Markovian processes with "memory" m

$$\frac{\mathsf{f}_1(X_n|X_{n-1},\dots,X_1)}{\mathsf{f}_0(X_n|X_{n-1},\dots,X_1)} = \frac{\mathsf{f}_1(X_n|X_{n-1},\dots,X_{n-m})}{\mathsf{f}_0(X_n|X_{n-1},\dots,X_{n-m})}$$

$$\frac{\mathsf{f}_1(X_n,\dots,X_1)}{\mathsf{f}_0(X_n,\dots,X_1)} = \frac{\mathsf{f}_1(X_n|X_{n-1},\dots,X_{n-m})}{\mathsf{f}_0(X_n|X_{n-1},\dots,X_{n-m})}\cdots\frac{\mathsf{f}_1(X_{m+1}|X_m,\dots,X_1)}{\mathsf{f}_0(X_{m+1}|X_m,\dots,X_1)} \times\frac{\mathsf{f}_1(X_m,\dots,X_1)}{\mathsf{f}_0(X_m,\dots,X_1)}$$

Can we estimate likelihood ratio of conditional densities?

a) Through data dynamics (classical)

b) Directly (proposed)

Classical Approach

Most common model, Autoregressive

$$X_t = A_1^i X_{t-1} + \cdots + A_m^i X_{t-m} + W_t, \quad i = 0, 1$$

$$X_t = G_i(X_{t-1}, \ldots, X_{t-m}, \theta^i) + W_t$$

Use training data $\{X_1^0, \ldots, X_{n_0}^0\}$ and $\{X_1^1, \ldots, X_{n_1}^1\}$ to solve

$$\min_{\theta^i} \sum_{t=1}^{n_i} \Big(X_t^i - G_i(X_{t-1}^i, \ldots, X_{t-m}^i, \theta^i) \Big)^2 \quad \Rightarrow \quad \theta_{\mathsf{o}}^i$$

$$W_t^i = X_t^i - G_i(X_{t-1}^i, \ldots, X_{t-m}^i, \theta_{\mathsf{o}}^i), \quad \Sigma_i = \frac{1}{n_i} \sum_{t=1}^{n_i} W_t^i (W_t^i)^\intercal$$

Assume $\{W_t^i\}$ i.i.d. Gaussian $\mathcal{N}(0, \Sigma_i)$

X_t given $\{X_{t-1}, \ldots, X_{t-m}\}$ under hypothesis H_i

Gaussian with mean $G_i(X_{t-1}, \ldots, X_{t-m}, \theta_0^i)$ and covariance Σ_i

To test $\{X_1, \ldots, X_n\}$

$$W_t^i = X_t - G_i(X_{t-1}, \ldots, X_{t-m}, \theta_0^i)$$

$$\frac{\mathrm{f}_1(X_t|X_{t-1}, \ldots, X_{t-m})}{\mathrm{f}_0(X_t|X_{t-1}, \ldots, X_{t-m})} = \frac{e^{-\frac{1}{2}(W_t^1)^\intercal \Sigma_1^{-1} W_t^1}}{e^{-\frac{1}{2}(W_t^0)^\intercal \Sigma_0^{-1} W_t^0}} \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}}$$

Not purely data driven

Gaussian assumption arbitrary, not necessarily suitable for all data!

16

Proposed Approach

$$\log\left(\frac{\mathsf{f}_1(X_t|X_{t-1},\ldots,X_{t-m})}{\mathsf{f}_0(X_t|X_{t-1},\ldots,X_{t-m})}\right) = \log\left(\frac{\frac{\mathsf{f}_1(X_t,X_{t-1},\ldots,X_{t-m})}{\mathsf{f}_1(X_{t-1},\ldots,X_{t-m})}}{\frac{\mathsf{f}_0(X_t,X_{t-1},\ldots,X_{t-m})}{\mathsf{f}_0(X_{t-1},\ldots,X_{t-m})}}\right) =$$

$$\log\left(\frac{\mathsf{f}_1(X_t,X_{t-1},\ldots,X_{t-m})}{\mathsf{f}_0(X_t,X_{t-1},\ldots,X_{t-m})}\right) - \log\left(\frac{\mathsf{f}_1(X_{t-1},\ldots,X_{t-m})}{\mathsf{f}_0(X_{t-1},\ldots,X_{t-m})}\right)$$

$$\mathsf{u}_{m+1}(X_t,\ldots,X_{t-m},\theta_{m+1}) \qquad \mathsf{u}_m(X_t,\ldots,X_{t-m+1},\theta_m)$$

$$\log\left(\frac{\mathsf{f}_1(X_t|X_{t-1},\ldots,X_{t-m})}{\mathsf{f}_0(X_t|X_{t-1},\ldots,X_{t-m})}\right) \approx$$

$$\mathsf{u}_{m+1}(X_t,\ldots,X_{t-m},\theta_{m+1}) - \mathsf{u}_m(X_{t-1},\ldots,X_{t-m},\theta_m)$$

Example: Testing Markov sequences (proof of concept)

Scalar observations $\{x_1, \ldots, x_n\}$

$w_t \sim \mathcal{N}(0,1)$, i.i.d.

$\mathrm{H}_0 : x_t = w_t$
$\mathrm{H}_1 : x_t = \mathrm{sign}(x_{t-1})\sqrt{|x_{t-1}|} + w_t$

$\mathrm{u}_2(x_t, x_{t-1}, \theta_2) : \; 2 \times 20 \times 1$
$\mathrm{u}_1(x_t, \theta_1) : \; 1 \times 10 \times 1$ (Exponential)
Training data $n_0 = n_1 = 100, 200, 500$

Testing $n = 20$, i.e. $\{x_1, \ldots, x_{20}\}$

$$\sum_{t=2}^{20} \mathrm{u}_2(x_t, x_{t-1}, \theta_2) - \sum_{t=2}^{19} \mathrm{u}_1(x_t, \theta_1) \underset{\mathrm{H}_0}{\overset{\mathrm{H}_1}{\gtrless}} \eta$$

100000×20 samples from H_0 and H_1