Lecture 13: 3/1/17

Generative vs. Discriminative

Disc. model is a conditional dist.

$$p(y \mid \vec{x}, \vec{\theta})$$

Gen. model is a joint dist.

$$p(\vec{x}, y \mid \vec{\theta}) = \underbrace{p(y \mid \vec{x})}_{\text{Disc. model}} \underbrace{p(\vec{x})}_{\text{Model of the Data instances}}$$

usually we write joint as $p(x|y)\,p(y)$

$\Rightarrow$ Gen. vs. Disc. tradeoff can be understood as choosing whether or not to model $p(\vec{x})$

$$p(\vec{x}) = \sum_y p(x|y)\,p(y)$$

$$p(y|\vec{x}) = \ldots \quad \text{by Bayes rule}$$

Bayes Classifier

← a.k.a. "Bayes Optimal Classifier"
← a.k.a. "Minimum Bayes Risk Decoder"

Two problems we care about:

① Density Estimation

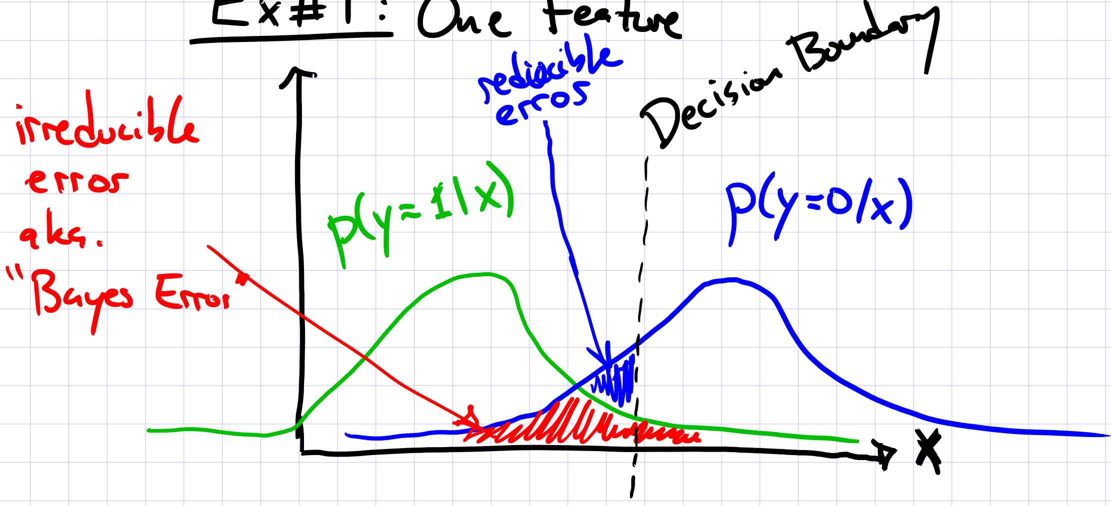
What does the distribution $p(x,y)$ look like?

② Choosing a Decision Function

How do we predict $\hat{y} = h(\vec{x})$? What is h?

Not the same problem!

Ex#1: One Feature



Assume:
- Instances $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$
- Given a probability distribution $p(x,y)$
- Given a loss function $\ell(y, y')$

  - Ex: 0-1 loss (for discrete y)

$$\ell(y,y') = \mathbb{1}(y \neq y') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{otherwise} \end{cases}$$

  - Ex: quadratic loss (for continuous y)

$$\ell(y,y') = (y-y')^2$$

Question: Given a new instance $\vec{x}$, what is the optimal prediction $\hat{y}$?

Answer:
$$\hat{y} = h(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(x,y) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(y|x)$$

a.k.a. "Bayes Risk"

Def: The expected loss risk(h) of a classifier $h(\vec{x})$ is:

a functional
a function of a function

$$\text{risk}(h) = E_{x,y \sim p(x,y)}\left[\ell(y, h(x))\right]$$
$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y)\, \ell(y, h(x))$$

Def: The <u>Bayes Classifier</u> h_{BC} is the h that minimizes the Bayes Risk, $\text{risk}(h)$

$$h_{BC} = \underset{h \in H}{\text{argmin}}\ \text{risk}(h)$$

Def: The <u>Bayes Error</u> is $\text{risk}(h_{BC})$

best we could possibly do.

Ex: Classification with 0/1 loss

Q: What is the Bayes Classifier?

$$\begin{aligned}
h_{BC} &= \underset{h}{\text{argmin}}\ \text{risk}(h) \\
&= \underset{h}{\text{argmin}} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y)\, \ell(y, h(x)) \\
&= \underset{h}{\text{argmin}} \sum_{x \in \mathcal{X}} \underbrace{\sum_{y \in \mathcal{Y}} p(x,y)\, \mathbb{I}(y \neq h(x))}_{G_x(h(x))} \\
&= \underset{h}{\text{argmin}} \sum_{x \in \mathcal{X}} G_x(h(x))
\end{aligned}$$

$\Rightarrow$ We want some h for each x, and it should return some $\hat{y}$ that minimizes $G_x(\hat{y})$

$$\begin{aligned}
h_{BC}(x) &= \underset{\hat{y}}{\text{argmin}}\ G_x(\hat{y}) \\
&= \underset{\hat{y}}{\text{argmin}} \sum_{y \in \mathcal{Y}} p(x,y)\, \mathbb{I}(y \neq \hat{y})
\end{aligned}$$

this term is 0 when $y = \hat{y}$

$$\begin{aligned}
&= \underset{\hat{y}}{\text{argmin}} \sum_{y \neq \hat{y}} \cancel{p(x,y)}\ p(y|x)\,p(x) \\
&= \underset{\hat{y}}{\text{argmin}}\ p(x) \sum_{y \neq \hat{y}} p(y|x) \\
&= \underset{\hat{y}}{\text{argmin}}\ p(x)\,(1 - p(\hat{y}|x)) \\
&= \underset{\hat{y}}{\text{argmin}}\ - \cancel{p(\hat{y}|x)\,p(x)}\ p(\hat{y},x) \\
&= \underset{\hat{y}}{\text{argmax}}\ p(\hat{y},x)
\end{aligned}$$

$\Rightarrow h(x) = \text{argmax}\ p(x,y)$ is the Bayes Classifier!

<u>Question</u>: Where does the distribution $p(x,y)$?

<u>Answer</u>: It's usually unknown.
So we try to learn it from data.

Maximum Likelihood Estimation

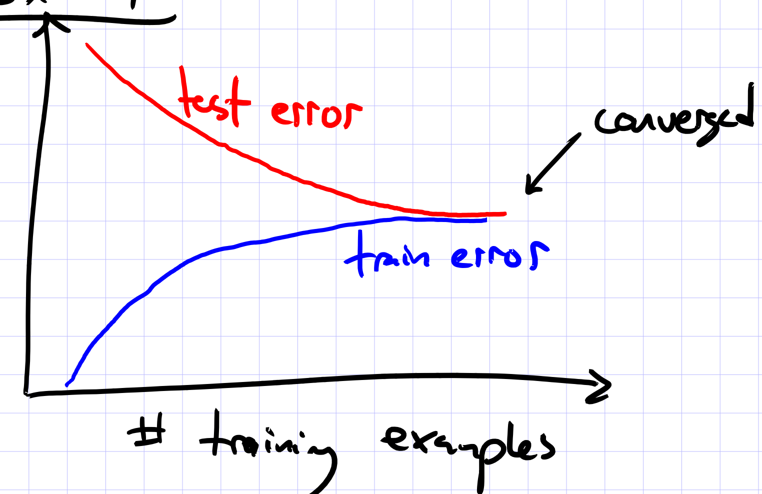
<u>Question</u>: Why should we use parameters that maximize likelihood?

<u>Answer</u>: Because the MLE θ^{MLE} is a <u>consistent</u> estimate of the true parameters θ^*

<u>Assume</u>: $x^{(i)}, y^{(i)} \sim p^*(x, y \mid \theta^*)$

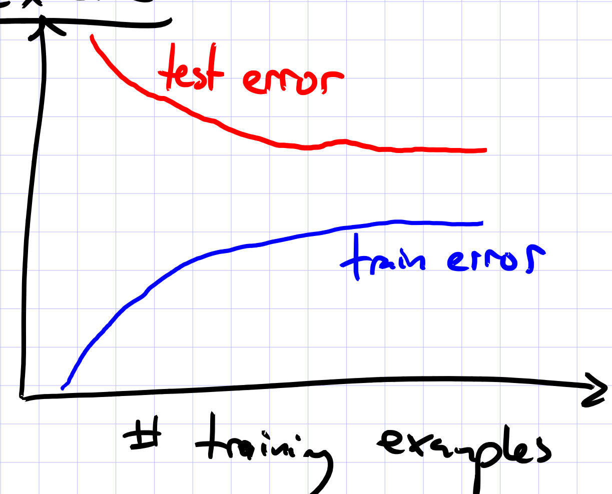
<u>Def</u>: A learning method is <u>consistent</u> if model error on new samples converges to model error on the original sample (as the size of original sample increases)

Ex: #1



<u>Consistent!</u>

Ex: #2



<u>Not Consistent!</u>

<u>Note</u>: The average likelihood converges almost surely to the expected log-likelihood by the strong law of large numbers.

$$\frac{1}{N} \sum_{i=1}^{N} \log p(x^{(i)}, y^{(i)}) \xrightarrow{a.s.} E_{x,y \sim p^*}[\log p(x,y)]$$

<u>Def</u>: <u>almost surely</u> convergence

$$\Pr\left(\lim_{n \to \infty} X_n = X\right) = 1$$

<u>Def</u>: Strong law of large numbers

$$\bar{X}_n \xrightarrow{a.s.} E[X]$$

as $n \to +\infty$

So not too surprising that given

$$\hat{\theta}^{(N)}_{MLE} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log p(x^{(i)}, y^{(i)} \mid \theta)$$

$$\theta^* = \text{true parameters}$$

We have that $\hat{\theta}^{(N)}_{MLE} \xrightarrow{a.s.} \theta^*$ (under some pairwise norm)

as $N \to \infty$

(Proof just requires KL divergence, and some prop)