Generative vs. Discriminative

Disc. model is a conditional dist. 
\( p(y|\tilde{x}, \theta) \)

Gen. model is a joint dist. 
\[ p(\tilde{x}, y|\theta) = p(y|\tilde{x}) p(\tilde{x}) \]

\[ p(y|\tilde{x}) = \sum_y p(x|y) p(y) \] 

\( p(\tilde{x}) \) by Bayes rule

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

\( \text{Disc. model \ Model as the} \) 
\( \text{Data instances} \)

\( \text{usually we write just} \) 
\( \text{as} \) 
\( p(x|y)p(y) \)
Bayes Classifier

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 \( \hat{y} = h(x) \)? What is \( h \)?

Not the same problem!

Ex#1: One Feature

\[ \text{Decision Boundary} \]

\[ p(y=1|x) \quad p(y=0|x) \]

Assume:
- Instances \( x \in X \) and labels \( y \in Y \)
- Given a probability distribution \( p(x, y) \)
- Given a loss function \( l(y, y') \)
  - Ex: 0-1 loss (for discrete \( y \))
    \[ l(y, y') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{otherwise} \end{cases} \]
  - Ex: Quadratic loss (for continuous \( y \))
    \[ l(y, y') = (y - y')^2 \]

Question: Given a new instance \( x' \), what is the optimal prediction \( \hat{y} \)?

Answer: \( \hat{y} = h(x) = \arg\max_{y \in Y} p(x, y) = \arg\max_{y \in Y} p(y|x) \)

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

\[ \text{risk}(h) = E_{x, y \sim p(x, y)} [l(y, h(x))] \]

\[ = \sum_{x \in X} \sum_{y \in Y} p(x, y) l(y, h(x)) \]
Def: The Bayes Classifier $h_{BC}$ is the $h$ that minimizes the Bayes Risk, $\text{risk}(h)$

$$h_{BC} = \arg\min_h \text{risk}(h)$$

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

We could possibly do

Ex: Classification with 0/1 loss

Q: What is the Bayes Classifier?

$$h_{BC} = \arg\min_h \text{risk}(h)$$

$$= \arg\min_h \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) l(y, h(x))$$

$$= \arg\min_h \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \mathbb{I}(y \neq h(x))$$

$$= \arg\min_h \sum_{x \in \mathcal{X}} G_x(h(x))$$

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

$$h_{BC}(x) = \arg\min_{\hat{y}} G_x(\hat{y})$$

$$= \arg\min_{\hat{y}} \sum_{y \in \mathcal{Y}} p(x,y) \mathbb{I}(y \neq \hat{y})$$

$$= \arg\min_{\hat{y}} \sum_{y \in \mathcal{Y}} p(x,y) p(y|x)$$

$$= \arg\min_{\hat{y}} p(x, \hat{y})$$

$$= \arg\min_{\hat{y}} p(\hat{y}|x) p(x)$$

$$= \arg\min_{\hat{y}} p(\hat{y}|x)$$

$$\Rightarrow h(x) = \arg\max_{\hat{y}} p(x, \hat{y})$$ is the Bayes Classifier!

Question: Where does the distribution $p(x,y)$?

Answer: It’s usually unknown.

So we try to learn it from data.
Maximum Likelihood Estimation

Question: Why should we use parameters that maximize likelihood?
Answer: Because the MLE \( \theta^{\text{MLE}} \) is a consistent estimate of the true parameter \( \theta^* \)

Assume: \( x^{(i)}, y^{(i)} \sim p^*(x, y | \theta^*) \)

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

Ex: #1

\[
\text{test error} \quad \xrightarrow{\text{converged}} \quad \text{train error}
\]

\# training examples

Consistent!

Ex: #2

\[
\text{test error}
\]

\# training examples

Not Consistent!

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

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

Def: almost surely convergence

\[
P_r(\lim_{n \to \infty} X_n = X) = 1
\]

Def: Strong law of large numbers

\[
X_n \overset{a.s.}{\to} E[X]
\]

as \( n \to \infty \)

So not too surprising that given

\[
\hat{\theta}^{(n)} = \arg \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log p(x^{(i)}, y^{(i)} | \theta)
\]

\( \theta^* = \) true parameters

We have that \( \hat{\theta}^{(n)} \overset{a.s.}{\to} \theta^* \) (under some premise now)

as \( N \to \infty \)

(Proof just requires KL divergence, and same prop)