Discriminative Models

Discriminative versus Generative

\[ y - \text{data}, \quad w - \text{state} \]

\[ P(w, y) \quad \Rightarrow \quad p(w | y) \]

\[ p(y | w) p(w) \quad \text{discriminative} \quad (\text{null space}) \]

Bayes Decision Theory.

Loss Function \( L(w, d(y)) \quad \text{d(y) decision} \)

Risk \[ R(d) = \sum_y L(w, d(y)) p(w | y) \]

Seek \( \hat{d} = \text{arg min}_d R(d) \quad \text{Bayes Decision Rule} \)

\( R(d) - \text{Bayes Risk} \)

Typically labelled examples:

\[ \frac{1}{N} \sum_{i=1}^{N} L(w_i, d(y)) \]

Empirical Risk:

\[ \text{Emp Risk}(d) = \sum_{i=1}^{N} L(w_i, d(y)) \]

Old-Fashioned Discriminative Rule.

Find \( d(y) \) directly.

\[ \text{linear separation} \]

Note: If the models are perfect, then there is little difference between discriminative & generative models.
**Discriminative Models**

\[
p(x|y, y) \quad \text{Graph } G = (S, E) \quad y - \text{data, } x - \text{labels} \\
\]

\[
x_i \rightarrow x_j \quad y_k \rightarrow y_i \\
\]

\[
p(x_i | y, x_{S \setminus y}) = p(x_i | y, x_{\text{N}(i)})
\]

N(i) - neighborhood of i

\[
p(x | y) = \frac{1}{Z} \exp \left( \sum_{i \in S} A_i(x_i, y) + \sum_{i \in S} \sum_{j \in \text{N}(i)} I_{ij}(x_i, x_j, y) \right)
\]

**Discriminative Markov Random Field (DMRF)**

neighborhood structure on x

\[A_i(x_i, y) \quad \text{association potential}
\]

\[I_{ij}(x_i, x_j, y) \quad \text{interactive potential}
\]

Alternative - generative models typically include factorized likelihood function:

\[
P(y | x) = \prod_i p(y_i | x_i)
\]

This is often unrealistic.

Building detectors: \[\square x_i \rightarrow \text{Building, No Building} \]

\[
y \]

\[ \]
Example:

\[ p(x_t = 1 | y) = \frac{1}{1 + e^{-(w_0 + w \cdot f_t(y))}} \]

model parameters \( w = (w_0, w) \) can be learned (e.g. AdaBoost)

\( f(y) \) denotes a set of features that can be computed from the image.

Usually assume that the features are independent of the position in the image (i.e. \( \text{indep of } t \))

Note: models of this type can be easily learnt from training data. They can be successful for detecting faces/text.
Discriminative Models

But $P(x_i = 1 | y)$ ignores spatial relationships.
The prob that $x_i$ is part of a building will often depend on $x_{n(i)}$, the neighborhood.

$$P_{\theta}(x_i | x_j, y) = \beta \left\{ K(x_i, x_j) + (1-\beta) \left( t_{ij} - \mu_j(y) - 1 \right) \right\}$$

Here $K(x_i, x_j)$ is an Ising model.
It is data independent $\Rightarrow$ encourages spatial stability.

Second term:
$$t_{ij} = \begin{cases} +1, & \text{if } x_i = x_j \\ -1, & \text{otherwise} \end{cases}$$

$\psi_j(\mathbf{y}, \mathbf{y}(y))$ is a feature vector.

This is like a smoothing term that adapts to the data.

The parameters for the association and interaction potentials can be learnt together.

$$p(x_i | y, \theta) \quad \theta = (\omega, \nu, \beta, K)$$

But the normalization function makes this difficult.
(5) Discriminative Models

pseudo-likelihood:

\[
\theta^m = \underset{\theta}{\arg \max} \prod_{m=1}^{M} \prod_{i \in s} P(x_i^m | x_n^m, Y^m, \theta)
\]

\[m\text{ index training image}\]

\[
P(x_i^m | x_n, y, \theta) = \frac{1}{Z_i} e^{-A(x_{i,y}) + \sum_{n \in i} I(x_n x_{i,y})}
\]

with \(Z_i = \sum_{x_i < (-L)} e^{-A(x_i, y)}\)

First learn the parameters for \(A(x_i, y)\) assuming independence \(\Rightarrow\) max/min likelihood regression.

Gives good initialization for constrained maxima/minima.

Inference

Given test image \(y\), goal is to find the optimal label configuration.

Algorithm:

for MAP estimation \(\Rightarrow\) Max-Flow/Min-cut.

provided \(K > 0.5\) and \(\beta > 0\)

True label.

Maximum Posterior Marginal (MPM)

\[
\text{cost function } C(x, x^*) = \sum_{i \in s} \left(1 - S(x, x_i^*)\right)
\]

\(\Rightarrow\) Belief Propagation or Sampling

Iterated Conditional Modes (ICM)

\[x_i < \underset{x_i}{\arg \max} P(x_i | X_n, y)\]