CSE 291: Domain Adaptation in Vision
Manmohan Chandraker

Lecture 6: Entropy
Papers for Wed, Feb 05

• Log-CORAL: Deep Domain Adaptation by Geodesic Distance Minimization
  • https://arxiv.org/abs/1707.09842

• Minimal-Entropy Correlation Alignment for Unsupervised Deep Domain Adaptation
  • https://arxiv.org/abs/1711.10288
Semi-Supervised Learning

• Goal: improve generalization using unlabeled data
  – In addition to supervision from labeled data

• Want to learn discriminative models
  – Generative modeling harder: \( p(x,y) \) is more complex than \( p(y|x) \)
  – Better models need not predict well (fitness not discriminative)

• Need ways to characterize information in unlabeled data
Information Content

• Consider an unfair coin with $p(\text{Heads}) = 0.99$
  – A coin toss that yields H is not surprising
  – But a toss that yields T is very surprising

• Information content of a stochastic event $E$
  \[ I(E) = -\log[Pr(E)] = -\log(P) \]
  – Logarithm in base 2: information in bits
  – Natural logarithm: information in nats

• For unfair coin,
  – Information in event Heads: $-\log(0.99) = 0.01$ bits
  – Information in event Tails: $-\log(0.01) = 6.64$ bits
  – Matches intuition for Tails being a more surprising event than Heads
Entropy

• Entropy: expected rate of information from stochastic process
  – For a random variable $X$, expected value is
    \[ E[X] = \sum_{i=1}^{n} x_i p_i \]
  – When the random variable is information, expectation is
    \[ H(X) = E[I(X)] = E[-\log(P(X))] = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \]
  – For the unfair coin with $p(H) = 0.99$, we have
    \[ H(X) = -(0.99 \log(0.99) + 0.01 \log(0.01)) = 0.08 \text{bits} \]

• More uncertainty = Higher entropy
  – The unfair coin delivers very little information (mostly just Heads)
  – For a fair coin with $p(\text{Heads}) = 0.5$, $H(X) = 1 \text{ bit}$
Conditional Entropy

- Conditional entropy of events $X$ and $Y$:

$$H(Y|X) = \sum_{x \in X} p(x)H(Y|X = x)$$

$$= - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log(p(y|x))$$

$$= - \sum_{x,y \in X,Y} p(x,y) \log(p(y|x))$$
KL Divergence

• Likelihood ratio
  – For a sample $x$, how much more likely to be from $p$ than from $q$
    \[ LR = \frac{p(x)}{q(x)} \]
  – If more likely from $p$, $LR > 1$, else $LR < 1$
  – Suppose several independent samples and consider log-ratios
    \[ LR = \sum_{i=0}^{n} \log \left( \frac{p(x_i)}{q(x_i)} \right) \]
  – Expected value of how much more likely is sampled data from $p$ than $q$
    \[ D_{KL}(P \parallel Q) = \sum_{i=0}^{n} p(x_i) \log \left( \frac{p(x_i)}{q(x_i)} \right) \]

• Information lost in expressing true distribution $P$, using $Q$
  \[ D_{KL}(P \parallel Q) = \sum_{i=0}^{n} p(x_i) \log(p(x_i)) - \sum_{i=0}^{n} p(x_i) \log(q(x_i)) \]

  Entropy  Cross-entropy
Semi-Supervised Learning

- Goal: improve generalization using unlabeled data
  - In addition to supervision from labeled data

- Want to learn discriminative models
  - Generative modeling harder: $p(x,y)$ is more complex than $p(y|x)$
  - Better models need not predict well (fitness not discriminative)

- Missing at random assumption
  - Missingness of labels is independent of the class information
  - Learning set: $\mathcal{L}_n = \{(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}, \ldots, x_n\}$
  - First $l$ examples labeled, next $u = n-l$ examples unlabeled
  - Encode missingness by random variable $h$
  - Let $h = 1$ when $y$ is missing and $h = 0$ when $y$ is observed
  - Then, we assume $P(h|x, y) = P(h|x)$
  - Situations when this is not a good assumption (“refuse to answer”)
Semi-Supervised Learning

• Information content of unlabeled examples
  – Not useful in an ML estimation which maximizes $\sum_{i=1}^{l} \ln P(y_i|x_i; \theta)$
  – Need to model the prior between $x$ and $y$
  – No universal prior, so seek an induction bias to process unlabeled data when they convey information
  – Information of unlabeled examples decreases when classes overlap [Castelli and Cover, 1996]

• Cluster (or low density separation) assumption
  – Better generalization if decision boundary lies in low density regions
  – Neighbors have similar activations, more likely for high density points to share labels
  – Make the most of unlabeled data when they are useful, that is, when the classes are far apart
Measure of Class Overlap

• Come up with a prior that makes $p(y|x)$ good for classification

• Conditional entropy
  – Related to usefulness of data only when labeling is ambiguous
  – Measure of class overlap should be conditioned on missingness

$$H(y|x, h = 1) = -\mathbb{E}_{xy} \left[ \ln P(y|x, h = 1) \right]$$

$$= -\int \sum_{m=1}^{M} \ln P(y = m|x, h = 1)p(x, y = m|h = 1) \, dx$$

• Expecting a low conditional entropy itself is not a unique prior
  – Can derive a prior by resorting to maximum entropy principle
  – Prior that verifies $\mathbb{E}_\theta \left[ H(y|x, h = 1) \right] = c$, where $c$ indicates how small conditional entropy should be on average:

$$p(\theta) \propto \exp \left( -\lambda H(y|x, h = 1) \right)$$
Maximum Entropy Principle

- Maximum entropy prior is best in absence of other information
  - Suppose testable information is known about a distribution
  - Expected value, some properties or constraints
  - Among all distributions that encode the information, find best one

- Consider a distribution with $k$ propositions
  - Most informative prior: if we know a specific proposition is always true
    - Entropy is 0
  - Least informative prior: if all propositions are equally true
    - Entropy is maximal value of $\log k$

- Choosing maximum entropy prior: pick least informative one
  - Otherwise our choice would be based on knowledge we do not possess
Maximum Entropy Principle

- Various constraints lead to known distributions as choices for maximum entropy prior
Measure of Class Overlap

• Empirical measure of conditional entropy
  – Replace expectation by the sample average
    \[ H_{\text{emp}}(y|x, h = 1; \mathcal{L}_n) = -\frac{1}{u} \sum_{i=l+1}^{n} \sum_{m=1}^{M} P(m|x_i, t_i = 1) \ln P(m|x_i, t_i = 1) \]
  – Missing at random assumption means \( P(y|x, h = 1) = P(y|x) \)
  – We have
    \[ H_{\text{emp}}(y|x, h = 1; \mathcal{L}_n) = -\frac{1}{u} \sum_{i=l+1}^{n} \sum_{m=1}^{M} P(m|x_i) \ln P(m|x_i) \]

• Equivalent to a prior that depends on values of missing data
Entropy Regularization

- MAP estimate: maximize the posterior distribution

\[ C(\theta, \lambda; \mathcal{L}_n) = L(\theta; \mathcal{L}_n) - \lambda H_{\text{emp}}(y|x, h = 1; \mathcal{L}_n) \]
\[ = \sum_{i=1}^{l} \ln P(y_i|x_i; \theta) + \lambda \sum_{i=l+1}^{n} \sum_{m=1}^{M} P(m|x_i; \theta) \ln P(m|x_i; \theta) \]

- Non-concave objective, contains local maxima
- Between a classification and a clustering objective
- Reflects intuition of low density separation in semi-supervised learning

Depends only on labeled data (Concave term)
Depends only on unlabeled data (Convex term)