EE/Stats 376A: Information Theory **Winter 2017**

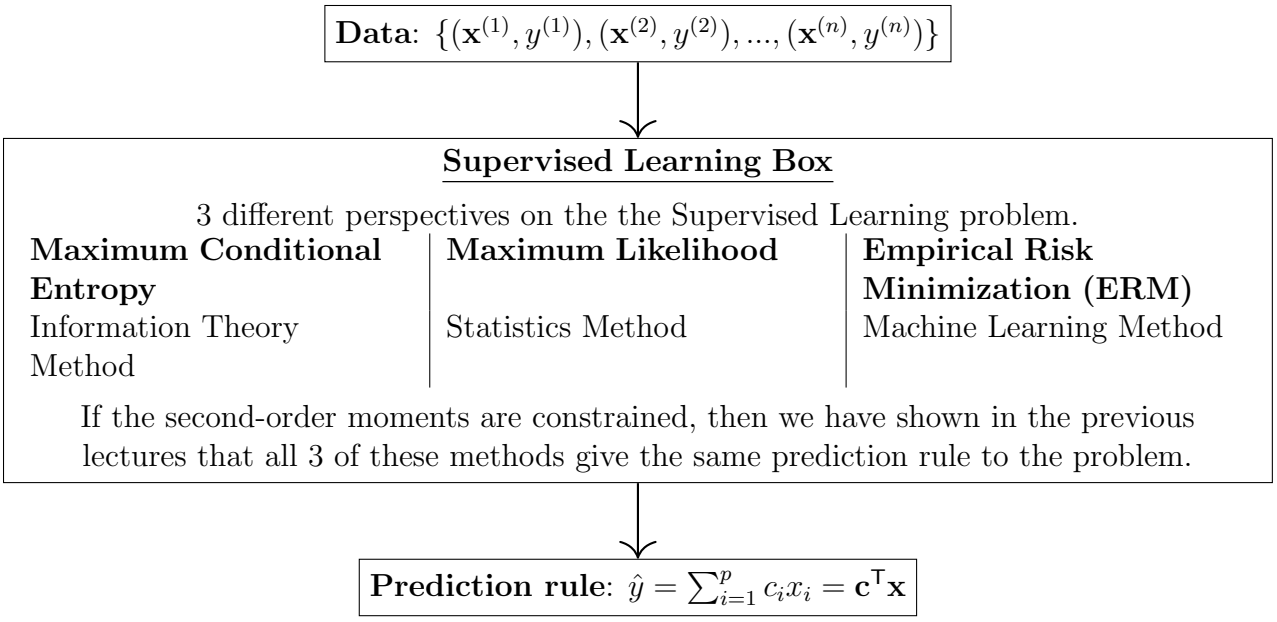
Lecture 19 — March 16

Lecturer: David Tse *Scribe: Yash Savani*

19.1 Outline

1. Information Theory, Machine Learning, and Statistics
2. Information vs Computational Limit

19.2 Overview



19.2.1 Maximum Conditional Entropy

The maximum conditional entropy method (for reference see [FT16] and [BPP96]) is the information theoretic perspective on supervised learning. In this method we maximize the conditional entropy of Y given $\mathbf{X}$ over a family of distributions Γ.

$$\max_{f \in \Gamma} H(Y|\mathbf{X}) \tag{19.1}$$

In the previous lectures, we saw that for second order moment constraints, the Gaussian distribution maximizes the conditional entropy (expression (19.1)). We also showed that

$H(Y|X)$ can be interpreted as:

$$\min_{\phi} \mathbb{E}[\ell_{\log}(Y, \phi(\mathbf{X}))]$$

with the optimal ϕ^* being $\phi^*(\mathbf{x}) = f(\cdot|\mathbf{x})$. Putting these two facts together enables us to obtain the solution (f^*, ϕ^*) to the following maxmin log loss problem:

$$\max_{f \in \Gamma} \min_{\phi} \mathbb{E}[\ell_{\log}(Y, \phi(\mathbf{X}))],$$

where $f^* = f_G$ being the jointly Gaussian distribution and $\phi^*(\mathbf{x}) = \phi_G(\mathbf{x})$ being the conditional distribution $f_G(\cdot|\mathbf{x})$. By a minimax theorem, the same (f_G, ϕ_G) is also the solution to the minimax problem:

$$\min_{\phi} \max_{f \in \Gamma} \mathbb{E}[\ell_{\log}(Y, \phi(\mathbf{X}))].$$

In particular, the minimax decision rule ϕ_G yields a conditional Gaussian distribution of Y given $\mathbf{x}$, and the conditional expectation of Y given $\mathbf{x}$ gives the linear prediction rule.

19.2.2 Maximum Likelihood

The maximum likelihood method is a systematic approach to statistical estimation and was proposed and popularized by Ronald Fisher in the 1910s. This method works by maximizing the likelihood of the data over a parameterization of the distribution the data is assumed to come from (for reference see [EH16]). We have mentioned that the maximum conditional entropy principle under second-order moment constraints is dual to the maximum likelihood problem under the Gaussian distribution:

$$\max_{\mathbf{c}, \sigma} \prod_{i=1}^{n} f_G(y^{(i)}|\mathbf{x}^{(\mathbf{i})}). \tag{19.2}$$

19.2.3 Empirical Risk Minimization (ERM)

The explicit solution of the maximum likelihood problem under the Gaussian model is:

$$\min_{\mathbf{c}} \sum_{i=1}^{n} (y^{(i)} - \mathbf{c}^{\mathsf{T}} \mathbf{x}^{(i)})^2 \tag{19.3}$$

This can be viewed as a special case of the empirical risk minimization approach used in machine learning (for reference see [Vap13]). This is an optimization problem over a class of predictors to produce one that best fits the data, as evaluated by a loss function. In this special case, the loss is squared error loss, and the predictor is linear in the input.

19.3 Generalization of the methods

If we want to change the loss function from squared error loss to 0-1 loss, a seemingly natural way to derive a new predictor is to modify the loss function used in the ERM method, since

it directly uses a loss function in the specification of the optimization problem. A possible ERM formulation for the 0-1 loss would look like

$$\min_{\mathbf{c}} \sum_{i=1}^{n} l_{0/1}(y^{(i)}, \text{sign}(\mathbf{c}^\mathsf{T}\mathbf{x}^{(i)})) \tag{19.4}$$

In words, if $\mathbf{c}^\mathsf{T}\mathbf{x}^{(i)}$ and $y^{(i)}$ have the same sign then the loss is 0, otherwise the loss is 1. Note: this is also known as the perceptron model.

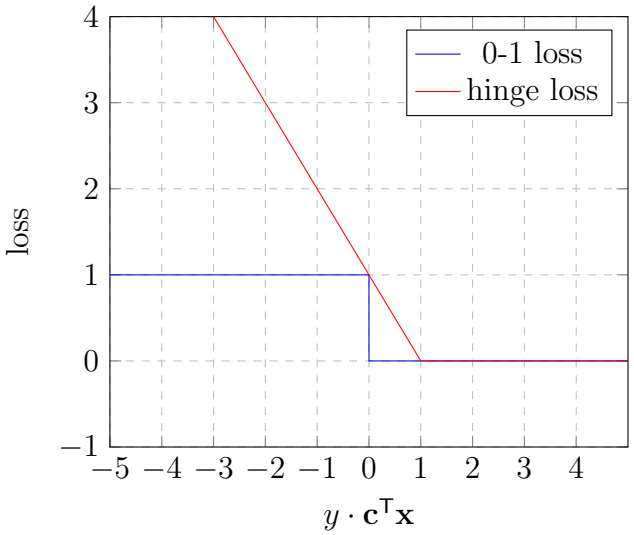


Figure 19.1: Plot of 0-1 loss and hinge loss functions

The **bad news** is that $l_{0/1}(y^{(i)}, \text{sign}(\mathbf{c}^\mathsf{T}\mathbf{x}^{(i)}))$ is a non-convex function in $\mathbf{c}$; we can see this from figure 20.1. It turns out that the corresponding optimization problem (perceptron problem) is NP-hard to solve optimally.

To solve this problem we usually relax the perceptron problem to get a convex loss. The closest convex loss function to the 0/1 loss function is the hinge loss. The solution to the ERM problem using the hinge loss is called a **Support Vector Machine (SVM)**. The SVM is the workhorse for many of the machine learning algorithms in use today. The hinge loss is referred to as the surrogate loss because it is a surrogate for the original 0-1 loss. The drawback is that this is only an approximation to the original $0-1$ loss problem.

An alternative approach we could take is to apply the maximum conditional entropy, or equivalently, the minmax approach, but to use $\ell_{0/1}$ instead of $\ell_{\log}$.

$$\min_{\phi} \max_{f \in \Gamma} \mathbb{E}[l_{0/1}(Y, \phi(\mathbf{X}))] \tag{19.5}$$

One key advantage of this approach is that no matter what the loss function is, the objective of this minimax problem is always linear in f and in ϕ. The potential drawback though is now f and ϕ are in the space of distributions, and therefore in very high dimensional space. If we can effectively reduce this dimensionality, we can solve the problem efficiently

using convex optimization methods. This turns out to be the case for $0-1$ loss and there is an efficient solution for the minimax $0-1$ loss. Moreover, the dual to this minimax $0-1$ loss problem turns out to be yet another ERM problem, not with a $0-1$ loss function but with a convex loss function called the minimax hinge loss. The plot of the minimax hinge loss is given in figure 19.2. In this formulation, we do not need to use a surrogate loss as in the ERM approach with $0-1$ loss.

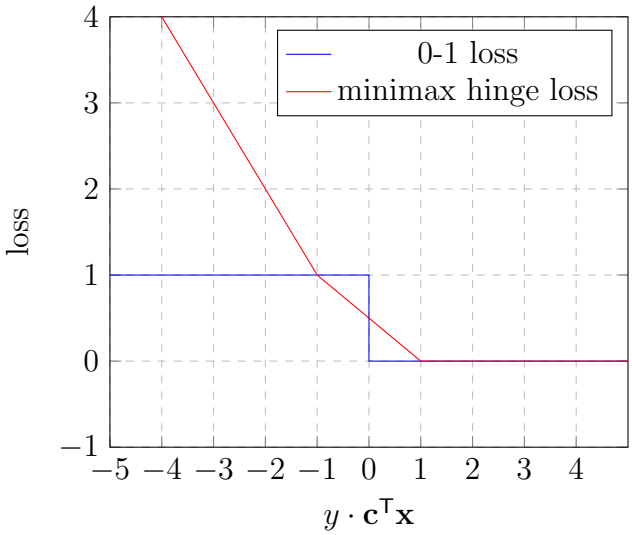


Figure 19.2: Plot of 0-1 loss and minimax hinge loss functions

.

19.4 Bibliographical Notes

The maximum entropy principle was first introduced by Jaynes [Jay57] in 1957. However, the game theoretic interpretation of this principle was discussed later in [Top79] and in [GD04] for a generalized defnition of entropy. The duality between the maximum conditional entropy problem and the maximum likelihood problem for logistic regression was shown in [BPP96]. The generalized version of this duality for a general loss function (including our discussion for the 0-1 loss case) is the subject of [FT16].

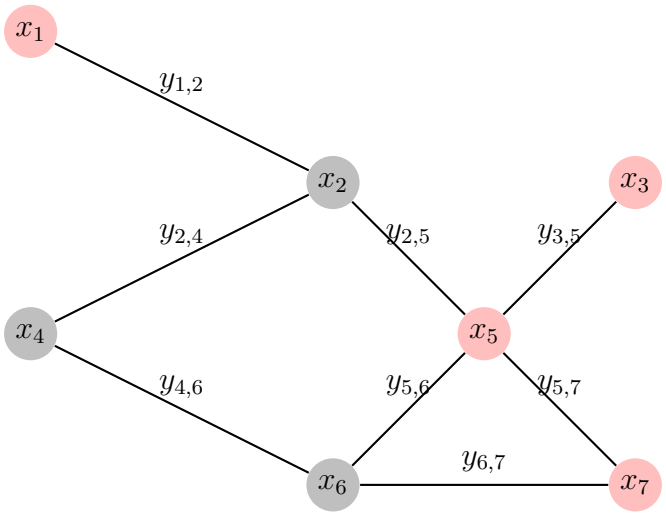
19.5 Themes of the course

In this course, there are two main themes. First we've defined important information measures such as Entropy (H), Mutual Information (I) and relative entropy. Second, we have defined the notion of *information limits* for specific problems such as compression and communication and showed that the limits can be attained efficiently. Moreover the information limits can be expressed in terms of the basic information measures. Now, we want to try and extend the notion of information limits to problems in structured machine learning such as community detection. It turns out that while for some problems, the information limit can be achieved efficiently, for others it is not know if this can be done.

19.6 Community Detection

Community detection is partitioning the vertices of a graph into communities (clusters) that are more densely connected. To be more precise, we have a graph $G = (V, E)$, with vertex set $V = \{1, 2, \cdots, k\}$, let the group of a vertex v_i be denoted by $x_i \in \{0, 1\}$. The 'value' of an edge between i, j is $y_{ij} = x_i \oplus x_j \oplus z_{ij}$, where $z_{ij} \sim Bern(p)$, and we observe n edges (out of $\frac{k(k-1)}{2}$ possible edges) uniformly at random.

Figure 19.3: Example of graph from community detection problem



Theorem 1. Let $X_i \overset{\text{iid}}{\sim} \text{Bern}(0.5)$, then for large enough k,

1. If $\frac{n}{k \log k} < c(p) \Rightarrow$ recovery is impossible, and
2. If $\frac{n}{k \log k} > c(p) \Rightarrow$ recovery is possible w.p. $\to 1$,

$$c(p) = \frac{1}{2(1 - 2^{-D(0.5||p)})},$$

where $D(0.5||p)$ is the relative entropy between Bern(0.5) and Bern(p).

Proof. Not part of the syllabus.

Notice that there is still a component missing. Even though we have shown that recovery is possible, we have not shown that this can be accomplished efficiently. Ideally, we want an algorithm that scales polynomially with k, the number of nodes, and turns out that we can achieve this result with an $O(k \log k)$ algorithm! Also, notice that in contrast to the communication problem, where the number of measurements (received symbols) n scales linearly with the number of information bits k, here the number of measurements need to scale with $k \log k$.

19.7 Hidden Clique Problem

Related to the community recovery problem is the hidden clique problem. A clique is a set of nodes which are fully connected i.e, every pair of vertices in clique is connected. Consider a graph $G = (V, E)$, where edge between any two vertices $u, v \in V$ is present with probability p and it has a clique of size k (the location of the planted clique is unknown).

Theorem 2. If the hidden clique of size $k \geq O(\log n)$ (where $|V| = n$), it can be recovered from the graph G.

The brute method is to check every $O(\log n)$-sized subset of vertices for a clique. This method requires $\binom{n}{\log n}$ steps, and hence inefficient. Unfortunately, nobody has yet found a faster method to find the hidden cliques of size $O(\log(n))$. In contrast to the community detection problem where the information iimit equaled the computational limit, the 'best' **efficient** algorithm to can only recover hidden cliques of size $O(\sqrt{n})$, much greater than $0(\log n)$.

Conjecture: Is there a **gap** between the information limit and the computational limit for the hidden clique problem?

This is a 40 year old conjecture!

Bibliography

[BPP96] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra, *A maximum entropy approach to natural language processing*, Computational linguistics **22** (1996), no. 1, 39–71.

[EH16] Bradley Efron and Trevor Hastie, *Computer age statistical inference*, Cambridge University Press, 2016.

[FT16] Farzan Farnia and David Tse, *A minimax approach to supervised learning*, Advances In Neural Information Processing Systems, 2016, pp. 4233–4241.

[GD04] Peter Grunwald and Philip Dawid, *Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory*, Annals of Statistics (2004), 1367–1433.

[Jay57] Edwin T Jaynes, *Information theory and statistical mechanics*, Physical review **106** (1957), no. 4, 620.

[Top79] Flemming Topsoe, *Information-theoretical optimization techniques*, Kybernetika **15** (1979), no. 1, 8–27.

[Vap13] Vladimir Vapnik, *The nature of statistical learning theory*, Springer, 2013.