Statistical Machine Learning II
Spring 2017, Learning Theory, Lecture 3

Jean Honorio jhonorio@purdue.edu

1 Information Theory

First, we provide some information theory background.

Definition 3.1 (Entropy)**.** *The entropy of a discrete random variable x of support $\mathcal{X}$ and probability mass function p is defined as:*

$$\mathbb{H}(x) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

A basic property of the entropy of a discrete random variable x is that:

$$0 \leq \mathbb{H}(x) \leq \log|\mathcal{X}|$$

In fact, the entropy is maximal for the discrete uniform distribution. That is, $(\forall x \in \mathcal{X})\ p(x) = 1/|\mathcal{X}|$, in which case $\mathbb{H}(x) = \log|\mathcal{X}|$.

Definition 3.2 (Conditional entropy)**.** *The conditional entropy of y given x is defined as:*

$$\begin{aligned}
\mathbb{H}(y|x) &= \sum_{v\in\mathcal{X}} p_x(v) H(y|x=v) \\
&= -\sum_{v\in\mathcal{X}} p_x(v) \sum_{y\in\mathcal{Y}} p_{y|x}(y|v)\log p_{y|x}(y|v) \\
&= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{xy}(x,y)\log p_{y|x}(y|x)
\end{aligned}$$

The conditional entropy can be expressed in terms of the entropy:

$$\mathbb{H}(y|x) = \mathbb{H}(x,y) - \mathbb{H}(x)$$

Definition 3.3 (Mutual information)**.**

$$\mathbb{I}(x,y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p_{xy}(x,y)\log\frac{p_{xy}(x,y)}{p_x(x)p_y(y)}$$

A basic property of the mutual information of random variables x and y is that:

$$\mathbb{I}(x, y) \geq 0$$

Furthermore, the mutual information can be expressed in terms of the entropy:

$$\mathbb{I}(x, y) = \mathbb{H}(x) - \mathbb{H}(x|y)$$

Note that random variables x and y are independent if and only if $\mathbb{I}(x, y) = 0$.

Definition 3.4 (Conditional mutual information)**.**

$$\mathbb{I}(x, y|z) = \mathbb{H}(x|z) - \mathbb{H}(x|y, z)$$

Definition 3.5 (Markov chain)**.** *Random variables x, y and z are said to form a Markov chain $x \to y \to z$ if and only if their joint probability distribution can be written as:*

$$p_{xyz}(x, y, z) = p_x(x) p_{y|x}(y|x) p_{z|y}(z|y)$$

Equivalently, random variables x, y and z are said to form a Markov chain $x \to y \to z$ if and only if x and z are conditionally independent given y, and thus $\mathbb{I}(x, z|y) = 0$.

2 Fano's inequality

Fano's inequality allows to provide *information-theoretic* lower bounds on the sample complexity. The setting for the analysis is as follows. Nature picks a "true" hypothesis $\overline{f}$ from some distribution of hypotheses. Then, a dataset S of n samples is produced, conditioned on the choice of $\overline{f}$. The learner then infers $\widehat{f}$ from the dataset S. The probability of error of the learner is given by $\mathbb{P}[\widehat{f} \neq \overline{f}]$. By lower-bounding this probability of error, one can find the necessary number of samples for learning. (Analyses as in the previous lecture allows to find a sufficient number of samples.)

Theorem 3.1 (Fano's inequality)**.** *For any random variable $\widehat{f}$ with k possible outcomes, such that $\overline{f} \to S \to \widehat{f}$, we have:*

$$\mathbb{P}[\widehat{f} \neq \overline{f}] \geq \frac{\mathbb{H}(\overline{f}|S) - \log 2}{\log k}$$

(See [1] if interested in the proof.)

Corollary 3.1 (Fano's inequality)**.** *For any random variable $\widehat{f}$ with k possible outcomes, such that $\overline{f} \to S \to \widehat{f}$, where $\overline{f}$ is chosen by nature uniformly at random (also from k possible outcomes), we have:*

$$\mathbb{P}[\widehat{f} \neq \overline{f}] \geq 1 - \frac{\mathbb{I}(\overline{f}, S) + \log 2}{\log k}$$

Proof. By property of the mutual information, we have $\mathbb{H}(\overline{f}|S) = \mathbb{H}(\overline{f}) - \mathbb{I}(\overline{f}, S)$. Since $\overline{f}$ is chosen uniformly at random from k possible outcomes, then $\mathbb{H}(\overline{f}) = \log k$ and we prove our claim. $\square$

The key in using Fano's inequality is to define a hypothesis class $\mathcal{F}$ for which $k = |\mathcal{F}|$ is large, while the mutual information $\mathbb{I}(\overline{f}, S)$ is small and of order n.

3 Upper Bounds on the Mutual Information

One key step in the application of Fano's inequality is to upper-bound the mutual information $\mathbb{I}(\overline{f}, S)$. Next, we revise some important definitions and inequalities from information theory.

Definition 3.6 (Kullback-Leibler (KL) divergence)**.** *Assume that a random variable x has support $\mathcal{X}$. Assume that there are two probability density functions p and q, which define two probability distributions $\mathcal{P} = p(\cdot)$ and $\mathcal{Q} = q(\cdot)$ respectively. The KL divergence is defined as:*

$$\mathbb{KL}(\mathcal{P}\|\mathcal{Q}) = \int_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

One important property of the KL divergence for independent random variables is the following. Let $\mathcal{P}_{xy} = p_{xy}(\cdot)$ and $\mathcal{P}_x\mathcal{P}_y = p_x(\cdot)p_y(\cdot)$. Assume that x and y are independent, and thus $\mathcal{P}_{xy} = \mathcal{P}_x\mathcal{P}_y$ and likewise, assume that $\mathcal{Q}_{xy} = \mathcal{Q}_x\mathcal{Q}_y$. We have:

$$\mathbb{KL}(\mathcal{P}_{xy}\|\mathcal{Q}_{xy}) = \mathbb{KL}(\mathcal{P}_x\|\mathcal{Q}_x) + \mathbb{KL}(\mathcal{P}_y\|\mathcal{Q}_y) \tag{1}$$

(The proof of the above might be left for homework very soon.)

Let $\mathcal{P}_{xy} = p_{xy}(\cdot)$ and $\mathcal{P}_x\mathcal{P}_y = p_x(\cdot)p_y(\cdot)$. We can define the mutual information as:

$$\begin{aligned}
\mathbb{I}(x, y) &= \mathbb{KL}(\mathcal{P}_{xy}\|\mathcal{P}_x\mathcal{P}_y) \\
&= \int_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{xy}(x, y) \log \frac{p_{xy}(x, y)}{p_x(x)p_y(y)} dxdy
\end{aligned}$$

By well-known identities $p_{\overline{f},S}(\overline{f}, S) = p_{\overline{f}}(\overline{f})p_{S|\overline{f}}(S)$ and $p_S(S) = \sum_{\overline{f}\in\mathcal{F}} p_{\overline{f},S}(\overline{f}, S)$, and since $\overline{f}$ follows a uniform distribution $p_{\overline{f}}(\overline{f}) = 1/k$, we have:

$$\begin{aligned}
\mathbb{I}(\overline{f}, S) &= \sum_{\overline{f}\in\mathcal{F}} \int_S p_{\overline{f},S}(\overline{f}, S) \log \frac{p_{\overline{f},S}(\overline{f}, S)}{p_{\overline{f}}(\overline{f})p_S(S)} dS \\
&= \sum_{\overline{f}\in\mathcal{F}} \int_S p_{\overline{f}}(\overline{f})p_{S|\overline{f}}(S) \log \frac{p_{\overline{f}}(\overline{f})p_{S|\overline{f}}(S)}{p_{\overline{f}}(\overline{f})p_S(S)} dS \\
&= \frac{1}{k} \sum_{\overline{f}\in\mathcal{F}} \int_S p_{S|\overline{f}}(S) \log \frac{p_{S|\overline{f}}(S)}{p_S(S)} dS
\end{aligned}$$

$$= \frac{1}{k} \sum_{\overline{f} \in \mathcal{F}} \mathbb{KL}(\mathcal{P}_{S|\overline{f}} \| \mathcal{P}_S)$$

In the above, we use the distribution $\mathcal{P}_{S|\overline{f}} = p_{S|\overline{f}}(\cdot)$ as well as the distribution $\mathcal{P}_S = p_S(S) = \frac{1}{k} \sum_{\overline{f} \in \mathcal{F}} p_{S|\overline{f}}(S)$.

Furthermore, from the convexity of the KL divergence, we can show that:

$$\mathbb{I}(\overline{f}, S) \leq \frac{1}{k^2} \sum_{f \in \mathcal{F}} \sum_{f' \in \mathcal{F}} \mathbb{KL}(\mathcal{P}_{S|f} \| \mathcal{P}_{S|f'}) \tag{2}$$

(The proof of the above might be left for homework very soon.)

4 Application: Empirical Risk Minimization with a Finite Hypothesis Class

Here we will prove a negative result in a setting similar to Theorem 2.1. First, some necessary definitions.

Definition 3.7. *The multivariate normal distribution of a random vector* $\mathbf{x} \in \mathbb{R}^k$ *with mean* $\boldsymbol{\mu} \in \mathbb{R}^k$ *and (symmetric and positive definite) covariance* $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ *is defined by the probability density function:*

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k \det \boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

For shortness, we write $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Let the distributions $\mathcal{N}_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, then:

$$\mathbb{KL}(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{2} \left(\mathrm{tr}({\boldsymbol{\Sigma}_2}^{-1} \boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{T}} {\boldsymbol{\Sigma}_2}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - k + \log \frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1} \right)$$

Note that when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$, the KL divergence becomes:

$$\mathbb{KL}(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{2} \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|^2 \tag{3}$$

(The proof of the above might be left for homework very soon.)

Next, we show our negative result. As we mentioned before, our main goal will be to upper-bound the mutual information $\mathbb{I}(\overline{f}, S)$ in order to apply Fano's inequality.

Theorem 3.2. *Assume that nature picks a "true" hypothesis* $\overline{f}$ *from some distribution of hypotheses with support* $\mathcal{F}$ *where* $|\mathcal{F}| = k$. *Then, a dataset of* n *samples is produced, conditioned on the choice of* $\overline{f}$. *The learner then infers* $\widehat{f}$ *from the dataset. Under the same setting as in Theorem 2.1, there exists a* *specific prediction problem and data distribution such that if* $n \leq \frac{\log k}{2} - \log 2$, *then learning fails, i.e.,*

$$\mathbb{P}[\widehat{f} \neq \overline{f}] \geq 1/2$$

for any mechanism (or algorithm) that a learner could use for picking $\widehat{f}$.

Proof. Recall that in Theorem 2.1, we assume that $\mathcal{F}$ is a finite set of hypotheses, i.e., $\mathcal{F} = \{f_1 \dots f_k\}$ where $k < +\infty$ and $(\forall j)\ f_j : \mathcal{X} \to \mathcal{Y}$.

Here, we further assume that $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{Y} = \{-1, +1\}$ and that $f_j(\mathbf{x})$ is the sign of the j-th element of the k-dimensional vector $\mathbf{x}$, i.e., $f_j(\mathbf{x}) = \mathrm{sgn}(x_j)$. (For clarity, we are now using a super-index for the sample index and a sub-index for the vector entry.) Assume that nature picks a "true" hypothesis $\overline{f}$ uniformly at random from $\mathcal{F}$. Then, a dataset $S = \mathbf{x}^{(1)}, y^{(1)} \dots \mathbf{x}^{(n)}, y^{(n)}$ of n samples is produced, conditioned on the choice of $\overline{f}$.

We assume that $\mathbb{P}[y = +1|\overline{f} = f_j] = \mathbb{P}[y = -1|\overline{f} = f_j] = 1/2$. We also assume that $\mathbf{x}|y = +1, \overline{f} = f_j \sim \mathcal{N}(\boldsymbol{\mu}^{(j)}, \mathbf{I})$ and $\mathbf{x}|y = -1, \overline{f} = f_j \sim \mathcal{N}(-\boldsymbol{\mu}^{(j)}, \mathbf{I})$ where $\mu_i^{(j)} = 1[i = j]$. That is, if $\overline{f} = f_j$ then every hypothesis $f \in \mathcal{F} - \{f_j\}$ has on average a 50% risk, since the sign of a normal random variable with mean zero is either $+1$ or -1 with 50% probability.

Note that for any pair of distributions $\mathcal{P}_{xy}$ and $\mathcal{Q}_{xy}$ where $p_y(+1) = p_y(-1) = q_y(+1) = q_y(-1) = 1/2$, we have:

$$\begin{aligned}
&\mathbb{KL}(\mathcal{P}_{xy} \| \mathcal{Q}_{xy}) = \sum_{y \in \{-1,+1\}} \int_{x \in \mathcal{X}} p_y(y) p_{x|y}(x) \log \frac{p_y(y) p_{x|y}(x)}{q_y(y) q_{x|y}(x)} dx \\
&= \frac{1}{2} \left(\int_{x \in \mathcal{X}} p_{x|y=+1}(x) \log \frac{p_{x|y=+1}(x)}{q_{x|y=+1}(x)} dx + \int_{x \in \mathcal{X}} p_{x|y=-1}(x) \log \frac{p_{x|y=-1}(x)}{q_{x|y=-1}(x)} dx \right) \\
&= \frac{1}{2} \left(\mathbb{KL}(\mathcal{P}_{x|y=+1} \| \mathcal{Q}_{x|y=+1}) + \mathbb{KL}(\mathcal{P}_{x|y=-1} \| \mathcal{Q}_{x|y=-1}) \right) \qquad (4)
\end{aligned}$$

By eq.(2), by eq.(1) since S is a dataset of n independent samples $(x^{(i)}, y^{(i)})$ for $i = 1 \dots n$, by eq.(4), and by eq.(3) since $x|y$ is normally distributed, we have:

$$\begin{aligned}
\mathbb{I}(\overline{f}, S) &\leq \frac{1}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{KL}(\mathcal{P}_{S|f_j} \| \mathcal{P}_{S|f_{j'}}) \\
&= \frac{n}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{KL}(\mathcal{P}_{x,y|f_j} \| \mathcal{P}_{x,y|f_{j'}}) \\
&= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left(\mathbb{KL}(\mathcal{P}_{x|f_j,y=+1} \| \mathcal{P}_{x|f_{j'},y=+1}) + \mathbb{KL}(\mathcal{P}_{x|f_j,y=-1} \| \mathcal{P}_{x|f_{j'},y=-1}) \right) \\
&= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left(\mathbb{KL}(\mathcal{N}(\boldsymbol{\mu}^{(j)}, \mathbf{I}) \| \mathcal{N}(\boldsymbol{\mu}^{(j')}, \mathbf{I})) + \mathbb{KL}(\mathcal{N}(-\boldsymbol{\mu}^{(j)}, \mathbf{I}) \| \mathcal{N}(-\boldsymbol{\mu}^{(j')}, \mathbf{I})) \right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \left(\frac{1}{2} \|\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(j')}\|^2 + \frac{1}{2} \|\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(j')}\|^2 \right) \\
&= \frac{n}{2k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} \|\boldsymbol{\mu}^{(j)} - \boldsymbol{\mu}^{(j')}\|^2 \\
&= \frac{n}{k^2} \sum_{j=1}^{k} \sum_{j'=1}^{k} 1[j \neq j'] \\
&= \frac{n(k^2 - k)}{k^2} \\
&\leq n
\end{aligned}
$$

By Corollary 3.1 and assuming a probability of error of at least $1/2$:

$$
\mathbb{P}[\widehat{f} \neq \overline{f}] \geq 1 - \frac{\mathbb{I}(\overline{f}, S) + \log 2}{\log k} \geq 1 - \frac{n + \log 2}{\log k} = \frac{1}{2}
$$

By solving for n in the above, we obtain that if $n \leq \dfrac{\log k}{2} - \log 2$, then we have that $\mathbb{P}[\widehat{f} \neq \overline{f}] \geq 1/2$. $\square$

References

[1] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley & Sons, 2nd edition, 2006.