Information Theory

Lecture 3: Applications to Machine Learning

Mark Reid

Research School of Computer Science
The Australian National University

2nd December, 2014

Outline

1. Prediction Error & Fano's Inequality
2. Online Learning
3. Exponential Families as Maximum Entropy Distributions

Loss and Bayes Risk

Machine learning is often framed in terms of *losses*. Given observations from $\mathcal{X}$ and predictors $\mathcal{A}$, a **loss function** $\ell : \mathcal{A} \to \mathbb{R}^{\mathcal{X}}$ assigns penalty $\ell_x(a)$ for predicting $a \in \mathcal{A}$ when $x \in \mathcal{X}$ is observed.

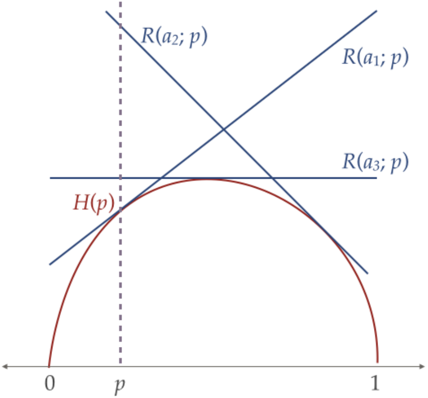
If observations come from fixed, unknown distribution $p(x)$ over $\mathcal{X}$ the **risk** of a is the expected loss

$$R(a; p) = \mathbb{E}_{x\sim p}\left[\ell_x(a)\right] = \langle p, \ell(a)\rangle$$

The **Bayes risk** is the minimal risk for any distribution

$$H(p) = \inf_{a\in\mathcal{A}} R(a; p).$$

$H(p)$ is always concave.



Log Loss and Entropy

In the special case when predictions are distributions over $\mathcal{X}$ (*i.e.*, $\mathcal{A} = \Delta_{\mathcal{X}}$) and the loss is **log loss**

$$\ell_x(q) = -\log q(x)$$

we get $R(q;p) = \mathbb{E}_{x\sim p}\left[-\log q(x)\right]$ and

$$H(p) = \inf_{q\in\Delta_{\mathcal{X}}} \mathbb{E}_{x\sim p}\left[-\log q(x)\right] = -\mathbb{E}_{x\sim p}\left[\log p(x)\right].$$

Furthermore, the **Regret** (*i.e.*, how far prediction was from optimal) is

$$\overbrace{R(q;p) - \inf_{q'\in\Delta_{\mathcal{X}}} R(q';p)}^{\mathrm{Regret}(q;p)} = \mathbb{E}_{x\sim p}\left[-\log q(x) + \log p(x)\right] = KL(p;q)$$

(Aside: In general, regret for a *proper* loss is always a *Bregman divergence* constructed from the negative Bayes risk of a loss)

Fano's Inequality

Fano's Inequality

Let $p(X, Y)$ be a joint distribution over X and Y where $Y \in \{1, \ldots, K\}$. If $\hat{Y} = f(X)$ is an estimator for Y then

$$p(\hat{Y} \neq Y) \geq \frac{H(Y|X) - 1}{\log_2 K}.$$

Proof: Define $E = 1$ if $\hat{Y} \neq Y$ and $E = 0$ if $\hat{Y} = Y$ and let $p = p(E = 1)$. Ignore X for the moment. Apply chain rule for conditional entropy:

$$H(E, Y|\hat{Y}) = H(Y|\hat{Y}) + H(E|Y, \hat{Y}) = H(E|\hat{Y}) + H(Y|E, \hat{Y})$$

- $H(E|Y, \hat{Y}) = 0$ since E is determined by Y and $\hat{Y}$.
- $H(E|\hat{Y}) \leq H(E) \leq 1$ (conditioning reduces entropy; E is binary)
- $H(Y|E, \hat{Y}) = (1 - p)\, H(Y|\hat{Y}, E = 0) + p\, H(Y|\hat{Y}, E = 1) \leq p \log_2 K$

since $E = 0 \implies Y = \hat{Y}$ and $H(Y|\hat{Y}) \leq H(Y) \leq \log_2 K$.

Fano's Inequality

Proof (cont.): So

$$H(Y|\hat{Y}) + 0 \leq 1 + p \log_2 K$$

But by the data processing inequality we know that $I(Y;\hat{Y}) \leq I(Y;X)$ since we assume $\hat{Y} = f(X)$ and so $Y \to X \to \hat{Y}$ forms a Markov chain. Thus,

$$I(Y;\hat{Y}) = H(Y) - H(Y|\hat{Y}) \leq H(Y) - H(Y|X) = I(Y;X)$$

which gives $H(Y|\hat{Y}) \geq H(Y|X)$ and so

$$H(Y|X) \leq 1 + p \log_2 K.$$

Rearranging gives Fano's inequality:

$$P(Y \neq \hat{Y}) \geq \frac{H(Y|X) - 1}{\log_2 K}$$

Fano's Inequality

We can interpret this inequality in some extreme situations to see if it makes sense.

$$P(Y \neq \hat{Y}) \geq \frac{H(Y|X) - 1}{\log_2 K}$$

Suppose we are trying to "learn noise". That is, that Y (the class label) is uniformly distributed and independ of X (the feature vector).

Then $H(Y|X) = H(Y) = \log_2 K$ and so Fano's inequality becomes:

$$P(Y \neq \hat{Y}) \geq \frac{\log_2 K - 1}{\log_2 K} = 1 - \frac{1}{\log_2 K}$$

Correct but weak since $P(Y \neq \hat{Y}) = 1 - \frac{1}{K}$ in this case.

Amount X tells us about Y bounds how well we can predict Y based on X.

Another Bound

We can also use obtain a bound on "chance matching".

Lower bound on match by chance

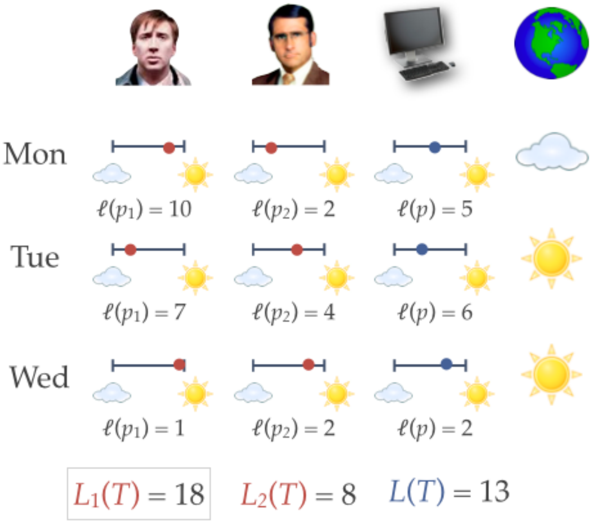
Suppose that Y and Y' are i.i.d. with distribution $p(Y)$. Then

$$p(Y = Y') \geq 2^{-H(Y)}.$$

This makes intuitive sense: the more "spread out" the distribution over Ys, the less chance we have of two randomly drawn samples matching. Conversely, if there is no randomness in Y then the probability of a match is 1.

Proof:
$p(Y = \hat{Y}) = \sum_y p(y)^2 = \mathbb{E}_{y\sim p}\left[2^{\log_2 p(y)}\right] \geq 2^{\mathbb{E}_{y\sim p}[\log_2 p(y)]} = 2^{-H(Y)}$.

Learning from Expert Advice: Motivation



Online Learning from Expert Advice

Consider the following game where each $\theta \in \Theta$ denotes an "expert" and $\ell : \Delta_{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is a loss.

Each round $t = 1, \ldots, T$:

1. Experts make predictions $p_\theta^t \in \Delta_{\mathcal{X}}$
2. Player makes prediction $p^t \in \Delta_{\mathcal{X}}$ (can depend on p_θ^t)
3. Observe a new instance $x^t \in \mathcal{X}$
4. Update losses: expert $L_\theta^t = L_\theta^{t-1} + \ell_{x^t}(p_\theta^t)$; player $L^t = L^{t-1} + \ell_{x^t}(p^t)$

Aim: choose p^t to minimise **regret** after T rounds $R(T) = L^T - \min_\theta L_\theta^T$

Ideally we want $R(T)$ so that $\lim_{T \to \infty} \frac{1}{T} R(T) = 0$ ("no regret").

No regret if $R(T) \propto \sqrt{T}$ ("slow rate") or if $R(T)$ is constant ("fast rate")

Mixable Losses and the Aggregating Algorithm

Vovk (1999) characterised when fast rates are possible in terms of a property of a loss he called mixability.

Mixable Loss

A loss $\ell : \Delta_{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is *η-mixable if for any $\{p_\theta \in \Delta_{\mathcal{X}}\}_{\theta \in \Theta}$ and any mixture $\mu \in \Delta_\Theta$ there exists $p \in \Delta_{\mathcal{X}}$ such that for all $x \in \mathcal{X}$*

$$\ell_x(p) \leq Mix_{\eta,\ell}(\mu, x) := -\frac{1}{\eta} \log \mathbb{E}_{\theta \sim \mu} [\exp(-\eta \ell_x(p_\theta))]$$

Examples:

- Log loss $\ell_x(p) = -\log p(x)$ is 1-mixable since $-\log \mathbb{E}_{\theta \sim \mu} [\exp(\log p_\theta(x))] = -\log \mathbb{E}_{\theta \sim \mu} [p_\theta(x)] = -\log p(x)$.
- Square loss $\ell_x(p) = \|p - \delta_x\|_2^2$ is 2-mixable.
- Absolute loss $\ell_x(p) = \|p - \delta_x\|_1$ is **not** mixable

Mixability Theorem

Mixability guarantees fast rates (*i.e.*, constant $R(T)$).

Mixability implies fast rates

If ℓ is an η-mixable loss then there exists an algorithm that acheive a regret

$$R(T) \leq \frac{\log|\Theta|}{\eta}.$$

The witness to the above result is called the **Aggregating Algorithm**:

- Initialise $\mu^0 = \frac{1}{|\Theta|}$
- Each round t
  - Set $\mu^t(\theta) \propto \mu^{t-1}(\theta) \exp(-\eta \ell_x(p_\theta))$
  - Predict using p guaranteed to satisfy $\ell_x(p) \leq \mathrm{Mix}_{\eta,\ell}(\mu, x)$

Proof of Mixability Theorem

When using AA to choose p^t, first note that if $W^t = \sum_\theta e^{-\eta L_\theta^t}$ then $\mathbb{E}_{\theta\sim\mu^t}\left[e^{-\eta \ell_{x^t}(p_\theta^t)}\right] = \sum_\theta e^{-\eta \ell_{x^t}(p_\theta^t)} e^{-\eta L^t}/W^t = W^{t+1}/W^t$.

Now consider total loss at round T when using AA to choose p^t:

$$
\begin{aligned}
L^T = \sum_{t=1}^{T} \ell_{x^t}(p^t) &\leq \sum_{t=1}^{T} \mathrm{Mix}_{\eta,\ell}(\mu^t, x^t) \\
&= \sum_{t=1}^{T} -\eta^{-1} \log \mathbb{E}_{\theta\sim\mu^t}\left[\exp(-\eta \ell_{x^t}(p_\theta^t))\right] \\
&= -\eta^{-1} \log \prod_{t=1}^{T} \frac{W^t}{W^{t-1}} = -\eta^{-1} \log \frac{W^T}{W^0} \\
&\leq \eta^{-1}\left(\eta L_\theta^T + \log|\Theta|\right)
\end{aligned}
$$

for all $\theta \in \Theta$, giving $L^T - L_\theta^T \leq \frac{\log|\Theta|}{\eta}$, as required.

What's this got to do with Information Theory?

The telescoping of W^t/W^{t-1} in the above argument can be obtained via an additive telescoping in the dual space to Δ_Θ since the mixability condition can be written as

$$\mathrm{Mix}_{\eta,\ell}(\mu, x) = \inf_{\mu' \in \Delta_\Theta} \overbrace{\mathbb{E}_{\theta\sim\Delta_\Theta}\left[\ell_x(p_\theta)\right]}^{\text{Fit the loss}} + \eta^{-1}\overbrace{KL(\mu'\|\mu)}^{\text{Regularise}}$$

and the minimising μ' is the distribution obtained from the AA.

Furthermore:

- The distributions $\mu^t(\theta) \propto e^{-\eta L^t_\theta}$ are like an EF with statistic $(L^t_\theta)_{\theta\in\Theta}$
- Regret bound is $\frac{1}{\eta}KL(\pi\|\delta_\theta)$ for $\pi = \frac{1}{N}$ and δ_θ is point mass on θ.
- For log loss, $\eta = 1$ and AA $=$ Bayesian updating

(Similar results hold for general Bregman divergence regularisation too)

Exponential Families: A Quick Review

Exponential Family

For *statistic* $\phi : X \to \mathbb{R}^d$ an exponential family (w.r.t. some measure λ) is a set $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ of densities of the form

$$p_\theta(x) := \exp\left(\langle \phi(x), \theta \rangle - C(\theta)\right)$$

with finite *cumulant* $C(\theta) := \log \int_X p_\theta(x)\, d\lambda(x)$. The parameters $\theta \in \Theta$ are natural parameters. The family $\mathcal{F}$ is regular if Θ is an open set

Exponential Families: A Quick Review

Exponential Family

For *statistic* $\phi : X \to \mathbb{R}^d$ an exponential family (w.r.t. some measure λ) is a set $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ of densities of the form

$$p_\theta(x) := \exp\left(\langle \phi(x), \theta \rangle - C(\theta)\right)$$

with finite *cumulant* $C(\theta) := \log \int_X p_\theta(x)\, d\lambda(x)$. The parameters $\theta \in \Theta$ are natural parameters. The family $\mathcal{F}$ is regular if Θ is an open set

Selected Properties:

- Convexity: Θ is a convex set. $C : \Theta \to \mathbb{R}$ is a convex function.
- The gradient of the cumulant is the mean: $\nabla C(\theta) = \mathbb{E}_{x \sim p_\theta}\left[\phi(x)\right]$
- The KL divergence $KL(p_\theta \| p_{\theta'}) = D_C(\theta', \theta)$ the BD for C

Exponential Families: A Quick Review

Exponential Family

For *statistic* $\phi : X \to \mathbb{R}^d$ an exponential family (w.r.t. some measure λ) is a set $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ of densities of the form

$$p_\theta(x) := \exp\left(\langle \phi(x), \theta \rangle - C(\theta)\right)$$

with finite *cumulant* $C(\theta) := \log \int_X p_\theta(x)\, d\lambda(x)$. The parameters $\theta \in \Theta$ are natural parameters. The family $\mathcal{F}$ is regular if Θ is an open set

Selected Properties:

- Convexity: Θ is a convex set. $C : \Theta \to \mathbb{R}$ is a convex function.
- The gradient of the cumulant is the mean: $\nabla C(\theta) = \mathbb{E}_{x \sim p_\theta}[\phi(x)]$
- The KL divergence $KL(p_\theta \| p_{\theta'}) = D_C(\theta', \theta)$ the BD for C

Exponential Families via Maximum Entropy

EF distributions are maximum entropy solutions with mean-constraints.

Maximum Entropy

Define the Shannon entropy $H(p) = -\int_X p(x) \log p(x) \, d\lambda(x)$. For a given mean value $r \in \mathbb{R}^d$ define the maximum entropy solution

$$p_r = \arg\sup\{H(p) : p \in \Delta_X, \mathbb{E}_p[\phi] = r\}$$

and the maximum entropy family $\mathcal{F} = \{p_r\}_{r \in \Phi}$.

Exponential Families via Maximum Entropy

EF distributions are maximum entropy solutions with mean-constraints.

Maximum Entropy

Define the Shannon entropy $H(p) = -\int_X p(x) \log p(x) \, d\lambda(x)$. For a given mean value $r \in \mathbb{R}^d$ define the maximum entropy solution

$$p_r = \arg\sup\{H(p) : p \in \Delta_X, \mathbb{E}_p[\phi] = r\}$$

and the maximum entropy family $\mathcal{F} = \{p_r\}_{r \in \Phi}$.

Properties:

- The exponential family $\{p_\theta\}_{\theta \in \Theta}$ and the MaxEnt family $\{p_r\}_{r \in \Phi}$ contain the same distributions
- A bijection between natural parameters $\theta \in \Theta$ and mean parameters $r \in \Phi$ is given by $r = \nabla C(\theta)$ and $\theta = (\nabla C)^{-1}(r) = \nabla C^*(r)$
- The Lagrangian $L(p, \theta) = H(p) + \langle \theta, \mathbb{E}_p[\phi] - r \rangle$ with dual vars θ.

Exponential Families via Convex Duality

Some simple calculations show that $-H$ is convex over Δ_X and

$$(-H)^*(q) = \log \int_X \exp\left(q(x)\right)\, d\lambda(x) \text{ and}$$

$$\nabla(-H^*)(q)_x = \frac{\exp(q(x))}{\int_X \exp\left(q(\xi)\right)\, d\lambda(\xi)}$$

Exponential Families via Convex Duality

Some simple calculations show that $-H$ is convex over Δ_X and

$$(-H)^*(q) = \log \int_X \exp\left(q(x)\right)\, d\lambda(x) \text{ and}$$

$$\nabla(-H^*)(q)_x = \frac{\exp(q(x))}{\int_X \exp\left(q(\xi)\right)\, d\lambda(\xi)}$$

Exponential Families via Convexity

For statistic $\phi : X \to \mathbb{R}^d$ each p_θ in the exp. family for ϕ can be written as

$$p_\theta = \nabla(-H)^*(\phi^\top \theta)$$

and $C(\theta) = (-H^*)(\phi^\top \theta)$ where $\phi^\top \theta \in \mathcal{W}^*$ denotes the RV $x \mapsto \langle \phi(x), \theta \rangle$.

Sufficient Statistics

Suppose we have a parametric family of distributions over $\mathcal{X}$, $\mathcal{P}_\Theta = \{p_\theta \in \Delta_\mathcal{X} : \theta \in \Theta\}$. In statistics, a sufficient statistic for intuitively captures all the information in observations from $\mathcal{X}$ for inference in $\mathcal{P}_\Theta$.

Sufficient Statistic

A function $\phi : \mathcal{X} \to \mathbb{R}^K$ for $\mathcal{P}_\Theta$ is a *sufficient statistic* if θ and X are conditionally independent given $\phi(X)$ — i.e., $\theta \to \phi(X) \to X$.

This intution can be formalised using mutual information via the data processing inequality: since $\phi(X)$ is a function of X we always have $\theta \to X \to \phi(X)$ and so $I(\theta; X) \geq I(\theta, \phi(X))$. However, DPI also say equality happens **iff** $\theta \to X \to \phi(X)$ — that is, iff ϕ is sufficient for $\mathcal{P}_\Theta$.

Sufficient Statistic (Info. Theory)

A function $\phi : \mathcal{X} \to \mathbb{R}^K$ is a *sufficient statistic* when $I(\theta; \phi(X)) = I(\theta; X)$.