Lecture 18 — March 14

Lecturer: David Tse *Scribe: Edwin Ng, Jack L, Vivek B*

18.1 Recap: Maximum Conditional Entropy Principle

The canonical problem of supervised learning is to find a good model for predicting Y from $\mathbf{X}$, where $\mathbf{X}$ is a high dimensional feature **vector**. If Y is continuous, this task is called *regression*, whereas if Y is discrete, this task is called *classification*. In many machine learning problems, the high dimensionality of $\mathbf{X}$ hampers us from determining the true distribution (model) of $\mathbf{X}, Y \sim p_{\mathbf{X},Y}$ from data. To overcome this limitation, in the previous lecture we proposed a conservative approach - choose a distribution from a set Γ with maximum conditional entropy i.e,

$$\operatorname*{argmax}_{P_{\mathbf{X},Y} \in \Gamma} H(Y \mid \mathbf{X}), \tag{18.1}$$

where the set Γ are derived from the data. For example, we could restrict ourselves to the distributions satisfying certain second-order moment constraints:

$$\Gamma = \{F_{\mathbf{X},Y} : \mathbb{E}[X_i Y] = \alpha_i, \mathbb{E}[X_j X_i] = \beta_{ij}, \mathbb{E}[Y^2] = \gamma\}, \tag{18.2}$$

where the constants $\beta_{ij}, \alpha_i, \gamma$ are estimated from the data.

18.2 Multivariate Gaussian

For the case when $Y \in \mathbb{R}$, we solve the the following optimization problem:

$$\begin{aligned} \operatorname*{argmax}_{f_{\mathbf{X},Y} \in \Gamma} \quad & h(Y \mid \mathbf{X}) \qquad \text{(MaxConEnt)} \\ s.t \quad & \Gamma = \{f_{\mathbf{X},Y} : \mathbb{E}[X_i Y] = \alpha_i, \mathbb{E}[X_j X_i] = \beta_{ij}, \mathbb{E}[Y^2] = \gamma\}. \end{aligned} \tag{18.3}$$

As in recent lectures, we will use the non-negativity property of the relative entropy to solve the above problem.

Let f^* be the solution of Problem 18.3, and let $f \in \Gamma$ be any arbitrary distribution. By the non-negativity property of the relative entropy, for every $\mathbf{x}$ we have

$$\begin{aligned} 0 &\le D\Big(f(\cdot \mid \mathbf{x}) \,||\, f^*(\cdot \mid \mathbf{x})\Big) \\ &\le -h(Y \mid \mathbf{X} = \mathbf{x}) + \mathbb{E}\left[\log \frac{1}{f^*(Y \mid \mathbf{X} = \mathbf{x})}\right] \qquad \text{where } (\mathbf{X}, Y) \sim f. \end{aligned}$$

Taking an expectation of the above equation over $\mathbf{X}$, we obtain

$$\mathbb{E}\left[\log \frac{1}{f^*(Y \mid \mathbf{X})}\right] \ge h(Y \mid \mathbf{X}) \qquad \text{where } (\mathbf{X}, Y) \sim f. \tag{18.4}$$

It is **sufficient** for the optimal distribution f^* to satisfy

$$\begin{aligned}\mathbb{E}\left[\log \frac{1}{f^*(Y \mid \mathbf{X})}\right] &= \mathbb{E}\left[\log \frac{1}{f^*(Y^* \mid \mathbf{X}^*)}\right] \qquad \text{where } (\mathbf{X}, Y) \sim f \text{ and } (\mathbf{X}^*, Y^*) \sim f^* \\ &= h(Y^* \mid \mathbf{X}^*), \end{aligned} \tag{18.5}$$

because it can be coupled with the inequality (18.4) to obtain

$$h(Y^* \mid \mathbf{X}^*) \geq h(Y \mid \mathbf{X}) \qquad \text{where } (\mathbf{X}, Y) \sim f \text{ and } (\mathbf{X}^*, Y^*) \sim f^*,$$

to prove the optimality of f^*.

Lemma 1. A distribution satisfying equation (18.5), f^*, is a Gaussian distribution i.e, $(\mathbf{X}_G, Y_G) \sim \mathcal{N}(0, K)$.

Proof. Since $(\mathbf{X}_G, Y_G)$ are jointly Gaussian, $Y_G \mid \mathbf{X}_G = \mathbf{x} \sim \mathcal{N}\left(\mathbf{c}^T\mathbf{x}, \sigma^2\right)$ for some vector $\mathbf{c} \in \mathbb{R}^p$ and σ^2, i.e.

$$f_G(y \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mathbf{c}^T\mathbf{x})^2/2\sigma^2},$$

and thus, we have

$$\mathbb{E}\left[\log \frac{1}{f_G(Y_G \mid \mathbf{X}_G)}\right] = \log\sqrt{2\pi\sigma^2} + \frac{\log e}{2\sigma^2}\mathbb{E}\left[(Y_G - \mathbf{c}^T\mathbf{X}_G)^2\right]. \tag{18.6}$$

The parameters $\mathbf{c}$ and σ^2 depend only on the second-order moments of $(\mathbf{X}_G, Y_G)$, which are fixed by the constraints in Γ (refer to equation 18.1). Therefore, for any $\mathbf{X}, Y \sim f \in \Gamma$,

$$\begin{aligned}\log\sqrt{2\pi\sigma^2} + \frac{\log e}{2\sigma^2}\mathbb{E}\left[(Y_G - \mathbf{c}^T\mathbf{X}_G)^2\right] &= \log\sqrt{2\pi\sigma^2} + \frac{\log e}{2\sigma^2}\mathbb{E}\left[(Y - c^T\mathbf{X})^2\right] \\ &= \mathbb{E}\left[\log \frac{1}{f_G(Y \mid \mathbf{X})}\right].\end{aligned}$$

Since f_G satisfies the condition (18.5), $f^* = f_G$. □

18.3 Connection to Maximum Likelihood

In the previous section, we showed that the optimal conditional distribution of Y given $\mathbf{X}$, $Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mathbf{c}^T\mathbf{x}, \sigma^2)$. The parameters c, σ^2 are some function (say h) of $(\alpha_i, \beta_{ij}, \gamma)$ (refer to equation (18.1)), which are themselves estimated using the data $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$:

$$\begin{aligned}\hat{\alpha}_i &= \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} y^{(k)} \\ \hat{\beta}_{ij} &= \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)} \\ \hat{\gamma} &= \frac{1}{n}\sum_{k=1}^{n} y^{(k)} y^{(k)} \\ \hat{c}, \hat{\sigma}^2 &= h(\hat{\alpha}_i, \hat{\beta}_{ij}, \hat{\gamma}).\end{aligned} \tag{18.7}$$

On the other hand, one could use maximum conditional entropy principle **only to motivate** the choice of the Gaussian distribution but **instead** use Maximum Likelihood (ML) to estimate $\mathbf{c}$, σ^2 from the data $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$ i.e,

$$\text{(Maximum Likelihood)} \qquad \mathbf{c}_{\text{ML}}, \sigma^2_{\text{ML}} = \operatorname*{argmax}_{\mathbf{c},\sigma^2} \prod_{k=1}^{n} f\left(y^{(k)} \mid \mathbf{x}^{(k)}; \sigma^2, \mathbf{c}\right). \tag{18.8}$$

For a given $\mathbf{X}$, the above predictor $f\big(y|\mathbf{x}; \mathbf{c}_{ML}, \sigma^2_{ML}\big)$ will return $y = \mathbf{c}_{ML}^T\mathbf{x}$, and hence solving the above ML ((18.8)) is solving *linear regression*.

By direct calculations, it turns out that both the methods give us the same answer. It turns out the deeper reason is that the ML problem happens to be the convex dual of the MaxConEnt problem.

18.4 Discrete Y: Logistic Regression

In the previous section we considered the case when $Y \in \mathbb{R}$. Let us now consider the case when $Y \in \{0, 1\}$, i.e. the binary classification problem. A natural maximum conditional entropy problem is

$$\begin{aligned} \operatorname*{argmax}_{F_{\mathbf{X},Y} \in \Gamma} \quad & h(Y \mid \mathbf{X}) \\ s.t \quad & \Gamma = \big\{F_{\mathbf{X},Y} : \mathbb{E}[X_i Y] = \alpha_i, \mathbb{E}[X_i X_j] = \beta_{ij}, Y \in \{0,1\}\big\}. \end{aligned} \tag{18.9}$$

The optimal solution to Problem (18.9) can be obtained using methods from the previous section, and it is equal to

$$f^*(y|\mathbf{x}) = g(x)e^{y\lambda^T\mathbf{x}},$$

for some function g, and some vector $\mathbf{c}$. Since $f^*(0|\mathbf{x})$ and $f^*(1|\mathbf{x})$ must sum up to 1, we have $g(\mathbf{x}) = \frac{1}{1+e^{\lambda^T\mathbf{x}}}$, and therefore

$$f^*(y|\mathbf{x}) = \frac{e^{y\lambda^T\mathbf{x}}}{1 + e^{\lambda^T\mathbf{x}}},$$

which is the logistic regression model!

18.5 Operational Interpretation

Finally, we turn to an operational interpretation for the maximum entropy principle. Generally for machine learning problems, we evaluate the performance of a prediction $\hat{y}$ by a *loss function* $\ell(y, \hat{y})$, where y is the true value. Examples

1. Mean-square loss: $\ell_2(y, \hat{y}) = (y - \hat{y})^2$.
2. Zero-one loss: $\ell_{0/1}(y, \hat{y}) = 1\{y \neq \hat{y}\}$.

The above two loss functions are defined for *hard predictors*; predictors which output a single value $\hat{y}$ for a given $\mathbf{X}$. In the case of *soft predictors*, the output $\hat{y}$ is an entire distribution q (over $\mathcal{Y}$), and a common choice of loss function is the ***log loss***

$$\ell(y, q) = \log \frac{1}{q(y)}.$$

Note that if $q(y)$ is small, the loss is large.

18.5.1 Unsupervised learning

In the case unsupervised learning with $\mathbf{Y} \sim p$, the optimal predictor for the *log loss* is

$$q^* = \operatorname*{argmin}_q \mathbb{E}\left[\ell(\mathbf{Y}, q)\right] = \underbrace{\mathbb{E}\left[\log \frac{1}{p(\mathbf{Y})}\right]}_{H(\mathbf{Y})} + D(p||q). \tag{18.10}$$

This quantity in expression (18.10) is called the *cross-entropy*, and we shall denote it by $C(p, q)$. For a given p $H(\mathbf{Y})$ is fixed, hence $q^* = p$ trivially. However, if we instead only known $p \in \Gamma$, a conservative approach would solve for the worst case scenario, and thus solve following minimax problem

$$\min_q \max_{p \in \Gamma} C(p, q).$$

It turns out that under some fairly mild assumptions (such as Γ convex and C a convex/concave function), we obtain

$$\begin{aligned} \min_q \max_{p \in \Gamma} C(p, q) &= \max_{p \in \Gamma} \min_q C(p, q) \\ &= \max_{p \in \Gamma} H(\mathbf{Y}). \end{aligned}$$

This is precisely the maximum entropy principle!

18.5.2 Supervised learning

In the supervised setting, the equivalent problem is to find a predictor $\phi : \mathbf{x} \mapsto q(\cdot \mid \mathbf{x})$, and thus the corresponding minimax problem turns out to be

$$\min_\phi \max_{p \in \Gamma} \mathbb{E}\left[\ell(Y, \phi(\mathbf{X}))\right]. \tag{18.11}$$

Similar to the case of unsupervised learning, under some fairly mild assumptions (such as Γ convex and C a convex/concave function), the 'min' and the 'max' in the expression (18.11) can be switched to obtain

$$\begin{aligned} \min_\phi \max_{p \in \Gamma} \mathbb{E}\left[\ell(Y, \phi(\mathbf{X}))\right] &= \max_{p \in \Gamma} \min_\phi \mathbb{E}\left[\ell(Y, \phi(\mathbf{X}))\right] \\ &= \max_{p \in \Gamma} H(Y \mid \mathbf{X}) \end{aligned}$$

which is precisely the maximum conditional entropy principle.