Machine Learning

Homework 2

Ari Feiglin

Exercise 2.1

(1) Provide a sufficient condition on the prior $\mathbb{P}(\theta)$ so that the MAP and MLE estimator coincide.

(2) Consider a dataset of n IID samples $x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma^2)$.

(i) Find the MLE estimates of the parameters μ and σ^2.

(ii) Assume that a prior distribution of μ is given by $\mathcal{N}(\mu_0, \sigma_0^2)$. Find the MAP estimator of the parameter μ.

(iii) What is the relation between the MLE and MAP estimators as $n \to \infty$ or $\sigma_0^2 \to \infty$.

(3) Find the MLE estimator of λ of a dataset $x_1, \dots, x_n \sim \mathrm{Poi}(\lambda)$.

(1) Since

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_\theta \left(\sum_i \log \mathbb{P}(x_i \mid \theta) + \log \mathbb{P}(\theta) \right), \qquad \hat{\theta}_{MLE} = \operatorname*{argmax}_\theta \sum_i \log \mathbb{P}(x_i \mid \theta)$$

It is sufficient that $\mathbb{P}(\theta)$ is constant over all possible θs, i.e. θ distributes uniformly. Then $\log \mathbb{P}(\theta)$ is constant and has no effect on the argmax.

(2)

(i) We know that

$$\widehat{\mu, \sigma^2}_{ML} = \operatorname*{argmax}_{\mu,\sigma} \sum_i \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) = \operatorname{argmax} \sum_i \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma) \right)$$

so let us take the gradient wrt (μ, σ),

$$\xrightarrow{\nabla} \sum_i \begin{pmatrix} \dfrac{x_i - \mu}{\sigma^2} \\ \dfrac{(x_i - \mu)^2}{\sigma^3} - \dfrac{1}{\sigma} \end{pmatrix}$$

Comparing with zero gives

$$\frac{1}{\hat{\sigma}^2} \sum_i (x_i - \hat{\mu}) = 0, \qquad \sum_i \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^2} - n = 0$$

The first equation can be simplified to $\sum_i (x_i - \hat{\mu}) = 0$, which gives $\hat{\mu} = \frac{1}{n} \sum_i x_i$. The second equation gives

$$\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2$$

(ii) The MAP estimator is the argmax of

$$\sum_i \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma) \right) - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \log(\sqrt{2\pi}\sigma_0)$$

Differentitiating wrt μ gives

$$\xrightarrow{\frac{\partial}{\partial \mu}} \sum_i \frac{x_i - \mu}{\sigma^2} - \frac{\mu - \mu_0}{\sigma_0^2}$$

Comparing to zero gives

$$\sigma_0^2 \sum_i x_i - \sigma_0^2 n\widehat{\mu} - \sigma^2\widehat{\mu} + \sigma^2\mu_0 = 0 \iff (n\sigma_0^2 + \sigma^2)\mu = \sigma^2\mu_0 + \sigma_0^2 \sum_i x_i$$

And so

$$\widehat{\mu} = \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_i x_i}{n\sigma_0^2 + \sigma^2}$$

(iii) When $n \to \infty$ or $\sigma_0^2 \to \infty$, $n\sigma_0^2 + \sigma^2 \sim n\sigma_0^2$ and so the MAP estimator acts like

$$\frac{\sigma^2\mu_0 + \sigma_0^2 \sum_i x_i}{n\sigma_0^2} = \frac{\sigma^2\mu_0}{n\sigma_0^2} + \frac{1}{n}\sum_i x_i \longrightarrow \frac{1}{n}\sum_i x_i$$

which is equal to the MLE estimator.

(3) The MLE estimator is the argmax of

$$\sum_i \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) = \sum_i (x_i \log\lambda - \lambda - \log(x_i!))$$

Differentiating wrt λ gives

$$\xrightarrow{\frac{\partial}{\partial\lambda}} \sum_i \frac{x_i}{\lambda} - 1$$

Comparing with zero gives

$$\widehat{\lambda} = \frac{1}{n}\sum_i x_i$$

Exercise 2.2

Let $\mathcal{H}$ be a hypothesis class which is PAC learnable whose sample complexity is given by $N(\varepsilon, \delta)$. Show that N is decreasing in both of its parameters.

This is obviously false. Our only restriction on N is that it exists and for every $n \geq N$, $\mathbb{P}(R(h_S) < \varepsilon \mid S \sim \mathcal{D}^n) > 1 - \delta$. So we can choose arbitrary ε, δ and increase $N(\varepsilon, \delta)$ to be arbitrarily large and it still satisfies the condition and is not decreasing. But perhaps you want that $\mathbb{P}(R(h_S) < \varepsilon \mid S \sim \mathcal{D}^n) > 1 - \delta \iff n \geq N$? In such a case we can prove it:

Let $0 < \varepsilon_1 < \varepsilon_2 < 1$ and $\delta \in (0, 1)$, then if we set $n = N(\varepsilon_1, \delta)$ we have that

$$\mathbb{P}(R(h_S) < \varepsilon_2 \mid S \sim \mathcal{D}^n) \geq \mathbb{P}(R(h_S) < \varepsilon_1 \mid S \sim \mathcal{D}^n) > 1 - \delta$$

So $N(\varepsilon_1, \delta) = n \geq N(\varepsilon_2, \delta)$

And let $0 < \delta_1 < \delta_2 < 1$ and $\varepsilon \in (0, 1)$, then if we set $n = N(\varepsilon_1, \delta_1)$ we have that

$$\mathbb{P}(R(h_S) < \varepsilon \mid S \sim \mathcal{D}^n) > 1 - \delta_1 > 1 - \delta_2$$

So $N(\varepsilon, \delta_1) = n \geq N(\varepsilon, \delta_2)$ as required.

Exercise 2.3

Given a real number $r \geq 0$, define the hypothesis $h_r \colon \mathbb{R}^d \longrightarrow \partial I$ by

$$h_r(x) = \begin{cases} 1 & \|x\| < r \\ 0 & \text{else} \end{cases}$$

Consider the hypothesis class $\mathcal{H} = \{h_r\}_{r \geq 0}$. Prove that it is PAC learnable in the realizable case. How does the sample complexity depend on d?

Let us define the algorithm $\mathcal{A}$ to get an input sample $S = (\vec{x}_1, \ldots, \vec{x}_n)$ and to simply return h_r where r is the largest norm of the $\vec{x}_i$s whose label is 1. This is indeed a PAC-learning algorithm for the problem, as we will prove. If we let $r = \max\|\vec{x}_i\|$ and h_R be the objective concept, then $R(h_S)$ is simply the probabilistic weight of the ring $\{\vec{x} \mid r < \|\vec{x}\| \leq R\}$. This would be when the probabilistic weight of the ring is at least ε (by uniformity), and so that we draw all of our samples from the area of $1 - \varepsilon$. This has a probability of occurring of $(1-\varepsilon)^n$, so we want

$$(1-\varepsilon)^n < \delta$$

Since $1 - \varepsilon < e^{-\varepsilon}$, we can require

$$e^{-\varepsilon n} < \delta \iff n > \frac{1}{\varepsilon} \log \frac{1}{\delta}$$

And so we can define the sample complexity to be $N(\varepsilon, \delta) = \frac{1}{\varepsilon} \log \frac{1}{\delta}$, and this satisfies the requirement. Notice that the sample complexity is independent of d.

Exercise 2.4

Call a hypothesis class $\mathcal{H}$ **PAC-learnable in expectation** if there exists an algorithm $\mathcal{A}$ and a function $N(a)\colon (0,1) \longrightarrow \mathbb{N}$ such that for all $a \in (0,1)$ and distribution $\mathcal{D}$, for every $n \geq N(a)$:

$$\mathbb{E}[R(h_S) \mid S \sim \mathcal{D}^n] \leq a$$

Show that $\mathcal{H}$ is PAC-learnable iff it is PAC-learnable in expectation.

Suppose $\mathcal{H}$ is PAC-learnable by $\mathcal{A}$, and let $N(\varepsilon, \delta)$ be its sample complexity. Then define $\tilde{N}(a) = N(a/2, a/2)$, and so for every $n \geq \tilde{N}(a)$:

$$\mathbb{E}[R(h_S)] = \mathbb{E}[R(h_S) \mid R(h_S) \leq a/2]\,\mathbb{P}(R(h_S) \leq a/2) + \mathbb{E}[R(h_S) \mid R(h_S) \geq a/2]\,\mathbb{P}(R(h_S) \geq a/2)$$

We know that $\mathbb{P}(R(h_S) \geq a) \leq a/2$ and since $R(h_S) \leq 1$ we have

$$\mathbb{E}[R(h_S)] \leq \frac{a}{2} \cdot 1 + 1 \cdot \frac{a}{2} = a$$

as required.

Now suppose $\mathcal{H}$ is PAC-learnable in expectation by $\mathcal{A}$. Then let $\tilde{N}(\varepsilon, \delta) = N(\varepsilon\delta)$, then for every $n \geq \tilde{N}(\varepsilon, \delta)$ we have that $\mathbb{E}[R(h_S)] \leq \varepsilon\delta$ and thus by Markov:

$$\mathbb{P}(R(h_S) \geq \varepsilon) \leq \frac{\mathbb{E}[R(h_S)]}{\varepsilon} \leq \frac{\varepsilon\delta}{\varepsilon} = \delta$$

as required. So $\mathcal{H}$ is PAC-learnable.