118

11/8/19

8805 Learning from Data: Lecture 22

- Followup on gaussian-processes / gaussian_processes_exercises.ipynb
  - Do sampling of different covariant functions in 2 Sampling from a Gaussian Process
    add: for i in range(nsamples):
              a.plot(X[:], Z[i,:])
    Predict nsamples=50 again.
  - Try some combinations
    - linear ⇒ polynomial (try two to get quadratic)
    see kernels.pdf Fig. 1.1 (pg. 2) and Fig. 1.2 (pg. 4)
  - For Gaussian Process Regression Model
    - distinguish between noise in data and noise in model
    - Compare the true function in red to the curves ⇐ now in official version

- Continue with Applications 1 and 2 on 116 and 117

- Other things to play with in gaussian_processes_exercises.ipynb
  - For GP Regression Model
    - Add a function for the true result (no noise) and add it (in red) to plots.
    - Try with small data noise and large data noise.
    - Try making lengthscale very small ⇒ explain the result.* (returns to mean after lengthscale)
    - Try optimizing with length scale very small → it doesn't change the lengthscale. ⇒ optimize fails.
      Fix by starting at a reasonable value.

    - With a good optimization, explore how well red line is within 95% bands.
      - relation to prior (mean zero, whatever variance).
      - What if I extend the range of Xtrue.

11/8/19

Maximum Entropy

Reference: Sivia, chapter 5 (on Carmen)

Notebooks: all in maximum-entropy on github under topics

- MaxEnt.ipynb
- demo-MaxEnt.ipynb
- Pdfs_from_MaxEnt.ipynb
- MaxEnt_Function_Reconstruction.ipynb

· Plan: First step through MaxEnt.ipynb
Then Pdfs_from_MaxEnt.ipynb as class exercise.
As time permits, do MaxEnt_Function_Reconstruction.ipynb.

MaxEnt.ipynb

- Ignorance pdfs → when we don't have constraints or extra knowledge that breaks symmetries.
  - i) permutation symmetry → die ⇒ 1/(# choices) (discrete)
  - ii) translational invariance → $p(x|I) =$ constant (in allowed region)
  - iii) scale invariance → $p(x|I) \propto \frac{1}{x}$
    - How to derive? First check that it works: $p(x|I) = \lambda p(\lambda x|I)$
      $\Rightarrow \frac{c}{x} = \frac{\lambda c}{\lambda x} = \frac{c}{x}$ ✓
      Now more general: $p(x|I) \propto x^\alpha \Rightarrow x^\alpha = \lambda(\lambda x)^\alpha = \lambda^{1+\alpha} x^\alpha$
      $\Rightarrow \alpha = -1$ ✓
      Still more general ⇒ set $\lambda = 1+\varepsilon$, with $\varepsilon \ll 1$, and solve to $O(\varepsilon)$: $p(x) = (1+\varepsilon)\left(p(x) + \varepsilon x \frac{dp}{dx}\right) \Rightarrow p(x) + x\frac{dp}{dx} = 0$
      $\Rightarrow \int_{p(x_0)}^{p(x)} \frac{dp}{p} = -\int_{x_0}^{x} \frac{dx}{x} \Rightarrow \ln\frac{p(x)}{p(x_0)} = \ln\frac{x_0}{x} \Rightarrow p(x) = \left(\frac{p(x_0)}{x_0}\right)\frac{1}{x}$ or $p(x) \propto \frac{1}{x}$ ✓

- Step quickly through Symmetry invariance
  - Basically using a change of variables for the symmetry, which means a Jacobian.

11/8/19

· For the linear model: $y_{th}(x) = \theta_1 x + \theta_0$, we could write it the other way around: $x_{th}(y) = \theta_1' y + \theta_0'$, and the probabilities should be equal:

$$p(\theta_0, \theta_1 | I)\, d\theta_0 d\theta_1 = p(\theta_0', \theta_1' | I)\, d\theta_0' d\theta_1'$$

· We can solve

$$y = \theta_1 x + \theta_0 = \theta_1(\theta_1' y + \theta_0') + \theta_0 = \theta_1 \theta_1' y + \theta_1 \theta_0' + \theta_0$$

$$\Rightarrow \theta_1 \theta_1' = 1, \quad \theta_1 \theta_0' + \theta_0 = 0 \Rightarrow \theta_1' = \theta_1^{-1}, \ \theta_0' = -\theta_1^{-1} \theta_0$$

· This lets us calculate the Jacobian: $\theta_1 = \theta_1'^{-1}$ $\theta_0 = -\theta_1 \theta_0' = -\theta_1'^{-1} \theta_0'$

$$\left| \frac{d\theta_0\, d\theta_1}{d\theta_0'\, d\theta_1'} \right| = \left| \det \begin{pmatrix} \frac{\partial \theta_0}{\partial \theta_0'} & \frac{\partial \theta_0}{\partial \theta_1'} \\ \frac{\partial \theta_1}{\partial \theta_0'} & \frac{\partial \theta_1}{\partial \theta_1'} \end{pmatrix} \right| = \begin{vmatrix} -\theta_1'^{-1} & \frac{\theta_0'}{\theta_1'^2} \\ 0 & -\theta_1'^{-2} \end{vmatrix} = \frac{1}{\theta_1'^3} = \theta_1^3$$

$$\Rightarrow p(\theta_0, \theta_1 | I)\, d\theta_0 d\theta_1 = p(-\theta_1^{-1}\theta_0, \theta_1^{-1} | I)\, d\theta_0 d\theta_1 \frac{1}{\theta_1^3}$$

or $\theta_1^3\, p(\theta_0, \theta_1 | I) = p(-\theta_1^{-1}\theta_0, \theta_1^{-1} | I)$

One possible solution is $p(\theta_0, \theta_1 | I) \propto (1 + \theta_1^2)^{-3/2}$

Principle of Maximum Entropy

· Arguing from monkeys distributing N balls in M boxes, so n_i in each box and $N = \sum_{i=1}^{M} n_i$.

· We'll let them do this many times, subject to constraints described by I.
  · The idea is to find the pdf specified by $p_i = n_i / N$ for all i that appears most often $\Rightarrow$ this best represents our state of knowledge.

· So this becomes a matter of counting microstates (ie a particular distribution $\{n_i\}$) that are most likely given the constraints.
  · We'll let $F(\{p_i\})$ = # ways to get $\{n_i\}$ / total # ways $= M^N$
  · Now do some combinatorics $\Rightarrow$ this is a multinomial distribution:

Stirling: $\log n! \approx n \log n - n$

$$\log F(\{p_i\}) = \log(N!) - \sum_{i=1}^{M} \log(n_i!) - N \log M \approx -N \log M + N \log N - \sum_{i=1}^{M} n_i \log n_i$$

$p_i = \frac{n_i}{N}$ $$\approx -N \log(M) - N \sum_{i=1}^{M} p_i \log p_i$$

11/8/19

- So the key piece to maximize is the entropy:

$$S = -\sum_{i=1}^{M} p_i \log(p_i)$$

- There are several arguments for maximizing the entropy:
  1) information theory: maximum entropy = minimum information (Shannon, 1948)
  2) logical consistency (Shore + Johnson, 1960)
  3) Uncorrelated assignments related monotonically to S (Skilling, 1988)

- The third one tells us that unless you know specifically about correlations, it should not be in your probability assignment. One finds that entropy maximization satisfies this condition (see the notebook for a comparison of different possibilities for a test problem)

- The continuous version of entropy is

$$S[p] = -\int p(x) \log\left(\frac{p(x)}{m(x)}\right)$$

where $m(x)$ is a measure function. It is there to ensure that $S[p]$ is invariant under $x \to y = f(x)$.
  - Typically this means $m(x) = \text{constant}$.

- Let's do the examples in Pdfs_from_MaxEnt.ipynb

Example 1: The Gaussian
constraints are normalization $\int_{-\infty}^{\infty} p(x)\,dx = 1$
and known variance: $\int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx = \sigma^2$

$$\Rightarrow \text{Maximize } Q(p; \lambda_0, \lambda_1) = -\int p(x) \ln \frac{p(x)}{m(x)}\,dx + \lambda_0\left(1 - \int p(x)dx\right) + \lambda_1\left(\sigma^2 - \int p(x)(x-\mu)^2 dx\right)$$

with uniform $m(x)$.

11/8/19

Step 1: $\frac{\delta Q}{\delta p(x)} = -\ln\frac{p(x)}{1} - \frac{p(x)}{p(x)} - \lambda_0 - \lambda_1 (x-\mu)^2$ (taking $m(x)=1$)

$$\frac{\partial Q}{\partial \lambda_0} = 1 - \int_{-\infty}^{\infty} p(x)\,dx \qquad \frac{\partial Q}{\partial \lambda_1} = \sigma^2 - \int_{-\infty}^{\infty} p(x)(x-\mu)^2\,dx$$

Step 2: $\frac{\delta Q}{\delta p(x)} = 0 \Rightarrow \ln p(x) = -(1+\lambda_0) - \lambda_1(x-\mu)^2$

$$\Rightarrow p(x) = e^{-(1+\lambda_0)}\, e^{-\lambda_1(x-\mu)^2}$$

Step 3: $\frac{\partial Q}{\partial \lambda_0} = 0 \Rightarrow \int_{-\infty}^{\infty} e^{-(1+\lambda_0)} e^{-\lambda_1(x-\mu)^2}\,dx = e^{-(1+\lambda_0)}\sqrt{\frac{\pi}{\lambda_1}} = 1 \Rightarrow e^{-(1+\lambda_0)} = \sqrt{\frac{\lambda_1}{\pi}}$

$$\frac{\partial Q}{\partial \lambda_1} = 0 \Rightarrow \int_{-\infty}^{\infty} e^{-(1+\lambda_0)}\, e^{-\lambda_1(x-\mu)^2}(x-\mu)^2\,dx = \underbrace{e^{-(1+\lambda_0)}\frac{1}{\lambda_1^{3/2}}}_{\frac{1}{\sqrt{\pi}\lambda_1}} \cdot \underbrace{\int_{-\infty}^{\infty} y^2 e^{-y^2}\,dy}_{\frac{\sqrt{\pi}}{2}} = \sigma^2$$

$$\frac{1}{\sqrt{\pi}\lambda_1} \cdot \frac{\sqrt{\pi}}{2} = \sigma^2 \Rightarrow \lambda_1 = \frac{1}{2\sigma^2}$$

$$\Rightarrow \boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}$$ Our friend the Gaussian

---

Example 2: The Poisson distribution

constraints are normalization $\int_0^\infty p(x)\,dx = 1$ $\quad x \geq 0$

and known mean: $\int_0^\infty x\,p(x)\,dx = \mu$

$\Rightarrow$ maximize $Q(p;\lambda_0,\lambda_1) = -\int p(x)\ln\left(\frac{p(x)}{m(x)}\right)dx + \lambda_0\left(1-\int p(x)\,dx\right) + \lambda_1\left(\mu - \int p(x)\,x\,dx\right)$
with uniform $m(x)$

So very similar: $\frac{\delta Q}{\delta p(x)} = -\ln\frac{p(x)}{1} - \frac{p(x)}{p(x)} - \lambda_0 - \lambda_1 x \Rightarrow \ln p(x) = -(1+\lambda_0) - \lambda_1 x$

$$\Rightarrow p(x) = e^{-(1+\lambda_0)}\, e^{-\lambda_1 x}$$

$e^{-(1+\lambda_0)}\int_0^\infty e^{-\lambda_1 x}\,dx = e^{-(1+\lambda_0)}\frac{1}{\lambda_1} = 1 \Rightarrow \lambda_1 = e^{-(1+\lambda_0)}$; $\int_0^\infty \underbrace{e^{-(1+\lambda_0)}}_{\lambda_1} \underbrace{e^{-\lambda_1 x} \cdot x\,dx}_{\frac{1}{\lambda_1^2}} = \mu \Rightarrow \lambda_1 = \frac{1}{\mu}$

$$\Rightarrow \boxed{p(x) = \frac{1}{\mu}e^{-x/\mu}}$$ Poisson distribution!

Try the other examples!