

Started on	Wednesday, 24 July 2024, 10:19 PM
State	Finished
Completed on	Wednesday, 24 July 2024, 10:53 PM
Time taken	33 mins 55 secs
Marks	42/42
Grade	10 out of 10 (100%)


Question 1

Correct

Mark 8 out of 8

One key ingredient to Bayesian learning is to model the stochastic process of interest as a joint distribution that depends on some to-be-learned parameters.

For the following examples, match the following descriptions of the modelling to the formula for the joint probability distribution.

- **Bayesian net of discrete** random variables with conditional probabilities $\theta_i : (x_i; p_1, \dots, p_n) \mapsto P(X_i = x_i \mid \text{Parents}(X_i) = p)$
- **Fully continuous Bayesian net** modeled as multivariate Gaussian, i.e., for X_i a linear model $x_i = g_{\theta_i}(p) = p^T \theta_{i,\omega} + \theta_{i,b}$ with normally distributed errors.
- **Linear regression** model $g_{\theta}(x) : \mathbb{R}^n \rightarrow \mathbb{R} = x^T \theta + \theta_0$ with $SSE(g_{\theta}; \mathbf{d}) =: \sigma^2$ and normally distributed errors
- **Deep neural network** $g_{\mathbf{W}, \mathbf{b}}$ with weights \mathbf{W} and biases \mathbf{b} , i.e., parameters $\theta = (\mathbf{W}, \mathbf{b})$; with errors normally distributed, but with variance depending on the proximity to the mean of the input training data.
- **Drawing of a binary property** X (e.g., "[student owns a sports car](#)" ) **with replacement** from a population, with probability $\theta = P(X = \text{true})$

$P(x, y \mid \theta) = P(y \mid x)P(x) = \square(g_{\theta}(x), \sigma(\|x - \bar{x}\|_2))(y)P(x)$

Deep neural network

$P(x, y \mid \theta) = P(y \mid x)P(x) = \square(g_{\theta}(x), \sigma)(y)P(x)$

Linear regression

$P(x_1, \dots, x_n \mid \theta) = \prod_i P(x_i \mid X_j = x_j \text{ for } X_j \in \text{Parents}(X_i))$
 $= \prod_i \theta_i(x_i; x_j \text{ for } X_j \in \text{Parents}(X_i))$

Discrete Bayesian net

$P(x_1, \dots, x_n \mid \theta) = \prod_i N(g_{\theta_i}((x_j)_{j \leq n; X_j \in \text{Parents}(X_i)}, \sigma_i)(x_i)$

Fully continuous Bayesian net

$\backslash(P(X=c\mid\theta) = \theta^c(1-\theta)^{1-c} \backslash)$

Drawing with replacement

Your answer is correct.

Correct

Marks for this submission: 8/8.

Question 2

Correct

Mark 6 out of 6

For Bayesian learning we start with defining a **model** for the joint probability distribution that depends on some parameters θ (see previous question). The model should allow us to calculate the **likelihood of parameters θ given some training data points \mathbf{d}** . This likelihood is $P(\mathbf{d}|\theta)$.

Once the model and some training data points \mathbf{d} are given, the goal of Bayesian learning is to calculate the **probability of new observations based on the information of the training data**: $P(X|\mathbf{d})$.

Thanks to Bayes' theorem and marginalization, this is proportional to a sum/integral over **observation probabilities conditioned on θ** , written $P(X|\theta)$, for all possible θ , each weighted by the **posterior probability of θ** given the training data $P(\theta|\mathbf{d})$.

Again, thanks to Bayes, the posterior of a θ rewrites to the product of the **likelihood $P(\mathbf{d}|\theta)$** of θ with the **prior $P(\theta)$** of θ .

$P(\mathbf{d}|\theta)$

$P(X|\theta)$

$P(\theta|\mathbf{d})$

$P(\theta|\mathbf{d})$

$P(\theta)$

$P(X|\mathbf{d})$

Your answer is correct.

Correct

Marks for this submission: 6/6.

Question 3

Correct

Mark 8 out of 8

The prior used in full Bayesian learning and in maximum a posteriori estimation allows to encode prior knowledge about the parameters into the calculation/optimization. In the following, some examples of prior knowledge are given that can be encoded using a respective prior. Map the descriptions to prior distribution.

- **Sparsification:** The parameters are sparse (L1 regularization)
- **Magnitude penalization:** The parameters have low values (L2 regularization)
- **Hard value limits (uniform):** The parameters are limited to the interval $[-1,1]$.
- **Discrete values:** The parameters can take one of a discrete set of predefined values.
- **Hard value limits (non-uniform):** The parameter must be in $[0,1]$, and most probably is 0.5.

θ (uniformly) distributed over $\{\theta_1, \dots, \theta_k\}$.

Discrete values

θ normally distributed around 0

Magnitude penalization

$P(\theta) = \text{Beta}(2,2)$

Hard value limits (non-uniform)

$P(\theta)$ very dense at 0, e.g., as in the Laplace distribution

Sparsification

$P(\theta) = \frac{1}{2} \text{ for } \theta \in [-1,1], \text{ else } 0$.

Hard value limits (uniform)

Your answer is correct.

Correct

Marks for this submission: 8/8.

Question 4

Correct

Mark 8 out of 8

Fill in the table.

Comparison of full BL, MAP, MLE, EM

	Full BL	MAP	MLE	EM
One needs to evaluate the outcome for any θ .	yes	no	no	no
One does have to pick a θ (point estimate).	no	yes	yes	yes
Involves optimization for θ .	no	yes	yes	yes
Allows to use information about the prior.	yes	yes	no	yes
Can be overly confident and biased.	no	no	yes	yes
Works for hidden variables.	yes	no	no	yes
Always requires an initial estimate of θ .	no	no	no	yes
Gets easily stuck in local maxima.	no	no	no	yes
Is a consistent estimate.	yes	yes	yes	no

Correct

Marks for this submission: 8/8.

Question 5

Correct

Mark 8 out of 8

Map the formulas for $P(X|\mathbf{d})$ to the visited Bayesian learning techniques for inference using a distribution parametrized by θ . Recall that we saw full Bayesian learning (Full BL), maximum a posterior estimation (MAP), maximum likelihood estimation (MLE), and expectation maximization (EM).

Here, D_h refers to hidden variable resp. d_h for their values, and d_o for observable variable values.

$P(X|\mathbf{d}) \approx P(X|\mathop{\mathrm{argmax}}_{\theta} \log P(\mathbf{d}|\theta))$

MLE

$P(X|\mathbf{d}) = \int_{\theta} P(X|\theta)P(\mathbf{d}|\theta)P(\theta)d\theta$

Full BL

$P(X|\mathbf{d}) \approx P(X|\lim_{t \rightarrow \infty} \mathop{\mathrm{argmax}}_{\theta} \mathbb{E}_{\mathbf{d}_h \sim P(\mathbf{D}_h|\mathbf{d}_o, \theta^t)} [\log P(\mathbf{d}_h, \mathbf{d}_o|\theta)P(\theta)])$

EM

$P(X|\mathbf{d}) \approx P(X|\mathop{\mathrm{argmax}}_{\theta} P(\mathbf{d}|\theta)P(\theta))$

MAP

Your answer is correct.

Correct

Marks for this submission: 8/8.

Question 6

Correct

Mark 4 out of 4

When can you apply the following simplifications?

- $P(\mathbf{d}|\theta) = P(d_1 \wedge \dots \wedge d_n|\theta) = \prod_i P(d_i|\theta)$: Applies if
 - ☐ data points independent
 - ☒ data points independent and identically distributed
 - ☐ data points identically distributed

Mark 1 out of 1

- $\arg\max_{\theta} P(\mathbf{d}|\theta) = \arg\max_{\theta} l(P(\mathbf{d}|\theta))$ holds for strictly monotonically increasing functions, such as $l(x) = \log(x)$
- $P(X|\mathbf{d}) \approx P(X|\theta_{\text{MAP}})$: Applies if
 - ☒ X conditionally independent of d given θ This is, e.g., not the case if the training data was not drawn independently from X (e.g., drawing without replacement)
 - ☐ θ has a uniform prior $P(\theta)$
 - ☐ θ_{MAP} maximizes $P(X|\theta)$

Mark 1 out of 1

- $P(X|\theta_{\text{MAP}}) \approx P(X|\theta_{\text{MLE}})$: Applies if
 - ☒ θ can be assumed to have a uniform prior
 - ☐ The prior of θ is unknown
 - ☐ θ can be assumed to have a normal prior

Mark 1 out of 1

Correct

Marks for this submission: 4/4.

[◀ 04. Quiz - Bayesian networks](#)

Jump to...

[06. Quiz - Linear Models ▶](#)