

- For every question, start your answer **on a new page**.
- Do not hand in your scratch paper. Please do hand in this exam sheet, or leave it behind on your table.
- You are not allowed to use books nor printed or handwritten formula sheets. See the final page for supplied formulas.

1. For each of the following questions, write an essential answer in **maximally 2 sentences**.

a. What's the difference between classification and regression?

Classification is supervised learning with discrete targets; regression has continuously valued targets.

b. What's the difference between classification and clustering?

Clustering is classification with unobserved targets.

c. What's more appropriate to say and why: "the likelihood of the parameters" or "the likelihood of the data"?

The likelihood function is a function of the parameters but not a function of the data, since the data is given. So say "the likelihood of the parameters".

d. Why does maximum likelihood estimation become a better approximation to Bayesian learning as you collect more data?

The likelihood tends to get 'narrower' and more informative than the prior, since the latter does not change with more data.

e. Consider a classification problem with Gaussian class-conditional sampling distributions  $p(x_n|\mathcal{C}_1)$  and  $p(x_n|\mathcal{C}_2)$  for the feature observations  $x_n$  and class priors  $p(\mathcal{C}_1) = \pi$ ,  $p(\mathcal{C}_2) = 1 - \pi$ . Under what condition (for the Gaussian class conditional distributions) is the discrimination boundary between the two classes a hyperplane?

Same variance:  $p(x_n|\mathcal{C}_1) = \mathcal{N}(\mu_1, \Sigma)$  and  $p(x_n|\mathcal{C}_2) = \mathcal{N}(\mu_2, \Sigma)$

f. Write out the sampling distribution  $p(x)$  for a Gaussian Mixture model with observations  $x$  and  $K$  hidden clusters.

$p(x) = \sum_k \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)$

g. Why is (probabilistic) principal components analysis more popular than factor analysis in the signal processing community?

In contrast to FA, PCA assumes the same variance for all input components, which is appropriate if the input components are shifted versions of each other. This occurs in the processing of time sequences.

h. Given is a model

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z}, \Psi)$$
$$p(\mathbf{z}) = \mathcal{N}(0, I)$$

Work out an expression for the marginal distribution  $p(\mathbf{x})$ .

$$p(\mathbf{x}) = \mathcal{N}(0, \mathbf{W}\mathbf{W}^T + \Psi)$$

- i. What's the difference between a hidden Markov model and a linear Dynamical system? For the same number of state variables, which of these two models has a larger memory capacity, and why?

HMM has binary-valued (on-off) states, where the LDS has continuously valued states. The latter holds more capacity because, eg, a 16-bit representation of a continuously-valued variable holds  $2^{16}$  different states.

- j. Consider a discriminative classification model where

$$p(C_k|x, \hat{\theta}) = \frac{\exp(\hat{\theta}_k^T x)}{\sum_{k'} \exp(\hat{\theta}_{k'}^T x)}$$

is the probability that feature  $x$  belongs to class  $k$ . Show that the discrimination boundary between two classes is a straight line.

The contour of equal probability is given by

$$\log \frac{p(C_k|x, \hat{\theta})}{p(C_j|x, \hat{\theta})} = (\hat{\theta}_k - \hat{\theta}_j)^T x \stackrel{!}{=} 0,$$

which is a line.

2. Consider a model  $p(x, z|\theta) = p(x|z, \theta)p(z|\theta)$  where  $x$  and  $z$  relate to observed and unobserved variables, respectively. Also available is an observed data set  $D = \{x_1, x_2, \dots, x_N\}$ . One iteration of the EM-algorithm for estimating the parameters  $\theta$  is described by

$$\hat{\theta}^{(m+1)} := \arg \max_{\theta} \left( \sum_z p(z|x = D, \hat{\theta}^{(m)}) \log p(x = D, z|\theta) \right).$$

- a. Apparently, in order to execute EM, we need to work out an expression for the 'responsibility'  $p(z|x = D, \hat{\theta}^{(m)})$ . Use Bayes rule to show how we can compute the responsibility that allows us to execute an EM step?

Use Bayes rule:

$$p(z|x = D, \hat{\theta}^{(m)}) = \frac{p(x = D|z, \hat{\theta}^{(m)}) p(z|\hat{\theta}^{(m)})}{\int p(x = D|z, \hat{\theta}^{(m)}) p(z|\hat{\theta}^{(m)}) dz}$$

- b. Why do we need multiple iterations in the EM algorithm?

We must have a parameter estimate in order to compute the responsibilities, and vice versa, we need responsibilities to update the parameter estimate. Thus, in the EM algorithm, we iterate between updating responsibilities (beliefs about  $z$ ) and parameter estimates (beliefs about  $\theta$ ).

- c. Why can't we just use simple maximum log-likelihood to estimate parameters, as described by

$$\hat{\theta} := \arg \max_{\theta} \log p(x = D, z|\theta)?$$

Because  $z$  is not observed.

3. Consider the following state-space model:

$$z_k = Az_{k-1} + w_k$$

$$x_k = Cz_k + v_k$$

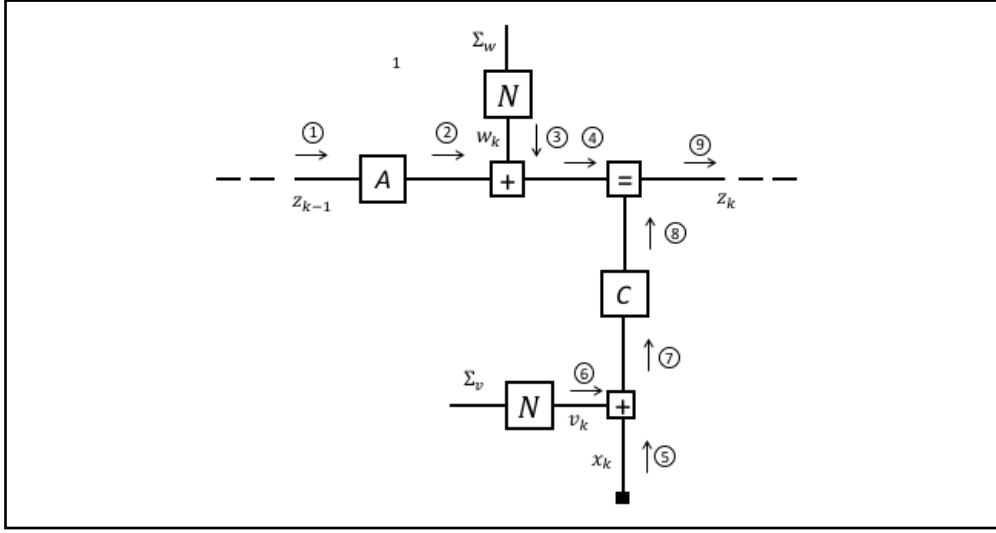
$$w_k \sim \mathcal{N}(0, \Sigma_w)$$

$$v_k \sim \mathcal{N}(0, \Sigma_v)$$

$$z_0 \sim \mathcal{N}(0, \Sigma_0)$$

where  $k = 1, 2, \dots, n$  is the time step counter;  $z_k$  is an *unobserved* state sequence;  $x_k$  is an *observed* sequence;  $w_k$  and  $v_k$  are (unobserved) state and observation noise sequences respectively;  $A$ ,  $C$ ,  $\Sigma_v$ ,  $\Sigma_w$  and  $\Sigma_0$  are known parameters.

- a. Draw the Forney-style factor graph for this model for the  $k^{th}$  time step.



- b. We are interested in estimating  $z_k$  from a given estimate for  $z_{k-1}$  and the current observation  $x_k$ , i.e., we are interested in computing  $p(z_k|z_{k-1}, x_k)$ . Can  $p(z_k|z_{k-1}, x_k)$  be expressed as a Gaussian distribution? Explain why or why not in one sentence.

(1 point). Yes, since the generative model  $p(x^n, z^n, \theta)$  is (one big) Gaussian.

- c. Draw the message passing schedule for computing  $p(z_k|z_{k-1}, x_k)$  by drawing arrows in the factor graph. Indicate the order of the messages by assigning numbers to the arrows ((1) for the first message; (2) for the second message and so on).

See the graph in box a.

4. A model  $\mathcal{M}$  is described (among others) by a single real valued parameter  $\theta$ ,  $0 < \theta < 1$  and a symbol alphabet  $\{0, 1, \dots\}$ .

The system described by  $\mathcal{M}$  can produce data  $d \in \{0, 1, \dots\}$  with a probability written as  $p(d|\mathcal{M}, \theta)$ . We also consider a prior distribution over  $\theta$  written as  $p(\theta|\mathcal{M})$ .

- a. Show that for any fixed data  $d$ ,  $p(\theta|\mathcal{M}, d)$  is a scaled version of  $p(\theta|\mathcal{M})p(d|\mathcal{M}, \theta)$ , or in other words, show that for any fixed  $d$ ,

$$\frac{p(\theta|\mathcal{M}, d)}{p(\theta|\mathcal{M})p(d|\mathcal{M}, \theta)}$$

does not depend on  $\theta$ .

From the chain rule we find

$$p(d, \theta | \mathcal{M}) = p(d | \mathcal{M}) p(\theta | \mathcal{M}, d) = p(\theta | \mathcal{M}) p(d | \mathcal{M}, \theta),$$

and so

$$\frac{p(\theta | \mathcal{M}) p(d | \mathcal{M}, \theta)}{p(\theta | \mathcal{M}, d)} = p(d | \mathcal{M}).$$

$p(d | \mathcal{M})$  does not depend on  $\theta$ , so we're done.

Let

$$\begin{aligned} p(\theta | \mathcal{M}) &= 6\theta(1 - \theta), \\ p(d | \mathcal{M}, \theta) &= (1 - \theta)\theta^d. \end{aligned}$$

- b. Determine the probability  $p(d | \mathcal{M})$  for the data  $d = 4$ .  
Note: compute the exact value.

$$\begin{aligned} p(d | \mathcal{M}) &= \int_0^1 p(\theta | \mathcal{M}) p(d | \mathcal{M}, \theta) d\theta, \\ &= \int_0^1 6\theta(1 - \theta) \cdot (1 - \theta)\theta^4 d\theta, \\ &= \int_0^1 6\theta(1 - \theta) \cdot (1 - \theta)\theta^4 d\theta, \\ &= \int_0^1 6\theta^5(1 - \theta)^2 d\theta, \\ &= \int_0^1 6\theta^5 - 12\theta^6 + 6\theta^7 d\theta, \\ &= \left. \frac{6}{6}\theta^6 - \frac{12}{7}\theta^7 + \frac{6}{8}\theta^8 \right|_0^1, \\ &= 1 - \frac{12}{7} + \frac{6}{8} = \frac{1}{28} = 0.0357. \end{aligned}$$

- c. Use the *Laplace approximation* for  $p(d | \mathcal{M})$  with the given  $p(\theta | \mathcal{M})$  and  $p(d | \mathcal{M}, \theta)$  for an estimated result.

Define  $f(\theta) = 6\theta^5(1 - \theta)^2$ . Then

$$\frac{\partial f(\theta)}{\partial \theta} = f'(\theta) = 6\theta(1 - \theta) \{5(1 - \theta) - 2\theta\}.$$

Solving for zero gives three solutions:  $\theta = 0$ ,  $\theta = \frac{5}{7}$ , and  $\theta = 1$ .

$f(\theta)$  has a maximum at  $\theta_0 = \frac{5}{7}$ .

Because  $f'(\theta_0) = 0$  we find

$$\left. \frac{\partial^2 \ln f(\theta)}{\partial \theta^2} \right|_{\theta=\theta_0} = \frac{f''(\theta_0)}{f(\theta_0)} = -\frac{343}{10}.$$

So our approximation is (with  $A = \frac{343}{10}$ )

$$\hat{p}(d | \mathcal{M}) = f(\theta_0) \sqrt{\frac{2\pi}{A}} = 0.039$$

5. We implement an e-mail spam filter using two features that we can extract from an e-mail. A feature can be the occurrence of a particular word or phrase in the e-mail.

Given an e-mail  $E$  we denote the extracted features by  $F$  and  $G$ .

$F = 1$  means that feature  $F$  is present in the e-mail  $E$ .

$F = 0$  means that feature  $F$  is absent. And likewise for feature  $G$ .

The variable  $C$  indicates whether  $E$  is spam ( $C = 1$ ) or not ( $C = 0$ ).

We are given 300 e-mails that are already classified. The following table shows how many e-mails contained certain features and the classification.

$F$	$G$	$C$	nr of e-mails
0	0	0	20
0	0	1	40
0	1	0	60
0	1	1	10
1	0	0	5
1	0	1	120
1	1	0	15
1	1	1	30

- a. From the table given above you can determine probability estimates using the maximum likelihood estimates. e.g. the probability  $P(C = 1)$ , i.e. the probability that an email will be spam, is approximated by:

$$P(C = 1) = \frac{\# \text{ of e-mails with } C = 1}{\text{total } \# \text{ of e-mails}} = \frac{200}{300} = 0.6667.$$

Note that the method using a beta prior would be better suited but we'll use the maximum likelihood because it is simpler.

Determine the following estimates.

$$\begin{aligned} &P(F = 1|C = 0), P(F = 1|C = 1), \\ &P(G = 1|C = 0), P(G = 1|C = 1), \\ &P(F = 0, G = 0|C = 0), P(F = 0, G = 1|C = 0), \\ &P(F = 1, G = 0|C = 0), P(F = 1, G = 1|C = 0), \\ &P(F = 0, G = 0|C = 1), P(F = 0, G = 1|C = 1), \\ &P(F = 1, G = 0|C = 1), P(F = 1, G = 1|C = 1). \end{aligned}$$

$P(F = 1 C = 0) = \frac{1}{5} = 0.2$	$P(F = 1 C = 1) = \frac{3}{4} = 0.75$
$P(G = 1 C = 0) = \frac{3}{4} = 0.75$	$P(G = 1 C = 1) = \frac{1}{5} = 0.2$
$P(F = 0, G = 0 C = 0) = \frac{1}{5} = 0.2$	$P(F = 0, G = 1 C = 0) = \frac{3}{5} = 0.6$
$P(F = 1, G = 0 C = 0) = \frac{1}{20} = 0.05$	$P(F = 1, G = 1 C = 0) = \frac{3}{20} = 0.15$
$P(F = 0, G = 0 C = 1) = \frac{1}{5} = 0.2$	$P(F = 0, G = 1 C = 1) = \frac{1}{20} = 0.05$
$P(F = 1, G = 0 C = 1) = \frac{3}{5} = 0.6$	$P(F = 1, G = 1 C = 1) = \frac{3}{20} = 0.15$

Model  $M_1$  for e-mail does not consider any feature. So  $P(C)$  can be used to estimate the probability that the next e-mail will be spam or not. We will write that as  $P(C|M_1)$ .

- b. Model  $M_2$  considers only feature  $F$  to predict whether the next e-mail will be spam or not. Use Bayes rule and the probability estimates determined in the previous question to determine an estimate for  $P(C|M_2) = P(C|F)$ .

Bayes	$P(C = 1 F = f) = \frac{P(C = 1)P(F = f C = 1)}{P(F = f)}, \quad f \in \{0, 1\}$
where	$P(F = f) = P(C = 0)P(F = f C = 0) + P(C = 1)P(F = f C = 1).$
We get	$P(C = 1 F = 0) = \frac{5}{13} = 0.3846$
and	$P(C = 1 F = 1) = \frac{15}{17} = 0.8824.$

Model  $M_3$  considers feature  $G$  only and model  $M_4$  considers both  $F$  and  $G$  and assumes that  $F$  and  $G$  are independent given the classification  $C$ .

- c. Use Bayes rule again to show how you would calculate  $P(C|M_4)$ .

Bayes	$P(C = 1 M_4) = \frac{P(C = 1)P(F = f C = 1)P(G = g C = 1)}{P(F = f, G = g)},$ $f \in \{0, 1\}; g \in \{0, 1\}$
where	$P(F = f, G = g) = P(C = 0)P(F = f C = 0)P(G = g C = 0) +$ $+ P(C = 1)P(F = f C = 1)P(G = g C = 1).$
We get	$P(C = 1 F = 0, G = 0) = \frac{2}{3} = 0.6667$
	$P(C = 1 F = 0, G = 1) = \frac{1}{7} = 0.1429$
	$P(C = 1 F = 1, G = 0) = \frac{24}{25} = 0.96$
and	$P(C = 1 F = 1, G = 1) = \frac{2}{3} = 0.6667.$

- d. The models  $M_1, M_2, \dots, M_4$  all have a certain number of free parameters. Determine the number of free parameters for each of the four models. HINT: Consider the number of free parameters of the joint distribution.

Model 1: No features so we consider the joint probability  $P\{C\}$  only. Thus we have one free parameter, for instance  $P\{C = 1\}$ .

Answer: 1 free parameter

Model 2: The joint probability is  $P\{C, F\} = P\{C\}P\{F|C\}$ . The free parameters are  $P\{C = 1\}$ ,  $P\{F = 1|C = 0\}$ ,  $P\{F = 1|C = 1\}$ .

Answer: 3 free parameters

Model 3: The joint probability is  $P\{C, G\} = P\{C\}P\{G|C\}$ . The free parameters are  $P\{C = 1\}$ ,  $P\{G = 1|C = 0\}$ ,  $P\{G = 1|C = 1\}$ .

Answer: 3 free parameters

Model 4: The joint probability is  $P\{C, F, G\} = P\{C\}P\{F|C\}P\{G|C\}$ . The free parameters are  $P\{C = 1\}$ ,  $P\{F = 1|C = 0\}$ ,  $P\{F = 1|C = 1\}$ ,  $P\{G = 1|C = 0\}$ ,  $P\{G = 1|C = 1\}$ .

Answer: 5 free parameters

- e. Given the training set of the 300 e-mail as shown in the table above, which of the five models would you prefer? Use an MDL argument in your answer.

HINT: You will need an estimate for the email entropy for each model. For model  $M_1$  you make an estimate of  $H(C)$  using the maximum likelihood estimate  $P(C = 1) = 0.6667$ . Likewise you calculate for  $M_2$  the entropy  $H(C|F)$  and thus you'll need to compute  $P(C, F)$ . For  $M_3$  you must compute the entropy  $H(C|G)$ ; for  $M_4$  you calculate  $H(C|F, G)$  and for  $M_5$  also  $H(C|F, G)$  although this will be a different calculation than for  $M_4$ .

The following table gives these estimated entropies and you may use them in your answer.

$$\begin{aligned} H(M_1) &= H(C) = 0.9183, & X \text{ is empty, so } P(C, X) &= P(C) \\ H(M_2) &= H(C|F) = 0.7127, & X = F, \text{ so } P(C, X) &= P(C)P(F|C) \\ H(M_3) &= H(C|G) = 0.7096, & X = G, \text{ so } P(C, X) &= P(C)P(G|C) \\ H(M_4) &= H(C|F, G) = 0.5604, & X = F, G, \text{ so } P(C, X) &= P(C)P(F|C)P(G|C) \end{aligned}$$

As Stochastic complexity we use the formula:

$$\text{sc} = \# \text{param} / 2 \log N + NH(C|\text{relevant params}),$$

Where  $N$  is the number of e-mails (300).

The stochastic complexities are:

$$\begin{aligned} \text{sc}(M_1) &= 1 \times \frac{1}{2} \log_2 N + NH(C) = 279.604 \\ \text{sc}(M_2) &= 3 \times \frac{1}{2} \log_2 N + NH(C|F) = 226.153 \\ \text{sc}(M_3) &= 3 \times \frac{1}{2} \log_2 N + NH(C|G) = 225.223 \\ \text{sc}(M_4) &= 5 \times \frac{1}{2} \log_2 N + NH(C|F, G) = 188.692 \end{aligned}$$

Based on this we prefer  $M_4$ , which has the lowest stochastic complexity.

## Appendix: formulas

$$\begin{aligned}|A^{-1}| &= |A|^{-1} \\ \nabla_A \log |A| &= (A^T)^{-1} = (A^{-1})^T \\ \text{Tr}[ABC] &= \text{Tr}[CAB] = \text{Tr}[BCA] \\ \nabla_A \text{Tr}[AB] &= \nabla_A \text{Tr}[BA] = B^T \\ \nabla_A \text{Tr}[ABA^T] &= A(B + B^T) \\ \nabla_x x^T A x &= (A + A^T)x \\ \nabla_X a^T X b &= \nabla_X \text{Tr}[ba^T X] = ab^T\end{aligned}$$

---

Points that can be scored per question:

- Question 1: all subquestions 1 point each. Total 10 points.
- Question 2: a) 2 points; b) 2 points; c) 1 point. Total 5 points.
- Question 3: a) 2 points; b) 1 point; c) 2 points. Total 5 points.
- Question 4: a) 2 points; b) 4 points; c) 4 points. Total 10 points.
- Question 5: a) 2 points; b) 2 points; c) 2 points; d) 2 points; e) 2 points. Total 10 points.

Max score that can be obtained: 40 points.

The final grade is obtained by dividing the score by 4 and rounding to the nearest integer.