Cribsheet for Machine Learning (COMS30035) Weeks 1–5

James Cussens

October 11, 2024

1 Machine Learning Principles

You need to know what the following terms mean:

- Unsupervised learning
- Supervised learning
- Regression
- Classification
- Underfitting
- Overfitting
- Model selection
- Training dataset
- Validation dataset
- Test dataset
- Cross-validation
- No free lunch theorem
- Model parameters
- Parametric model
- Nonparametric model
- Likelihood function
- Maximum likelihood estimation (MLE)

2 Linear Regression

You need to know that the linear regression model is:

$$p(y|\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \mathbf{x} + \epsilon \tag{1}$$

where ϵ has a Gaussian distribution with mean 0: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. You need to know what a *bias* parameter is and how in (1) it was included in the parameter vector $\mathbf{w}$ by the addition of a 'dummy' variable which always has the value 1.

You need to understand that we can apply linear regression to a *feature vector* $\phi(\mathbf{x})$ rather than the original data $\mathbf{x}$:

$$p(y|\phi(\mathbf{x}), \mathbf{w}) = \mathbf{w}^\top \phi(\mathbf{x}) + \epsilon \tag{2}$$

You need to know:

1. what the least-squares problem for linear regression is
2. that the solution to this problem has a closed-form (but you don't need to memorise this closed-form)
3. and that the least-squares solution is also the maximum likelihood solution (you do not need to be able to prove this).

3 Linear Discriminant

You need to know that when there are two classes Linear Discriminant computes $y = \mathbf{w}^\top \mathbf{x}$ for input x and assigns x to class C_1 if $y \geq 0$ and class C_2 otherwise. Parameters are 'learnt' by assuming that: (1) data for each class have a Gaussian distribution, (2) these 2 Gaussian distributions have the same covariance matrix. Parameters can then be found by applying MLE.

4 Logistic Regression

You need to know that the *logistic sigmoid function* (sometimes called just the *logistic function*) is:

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

You need to know that the logistic regression model for two classes is:

$$p(C_1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) \qquad p(C_2|\mathbf{x}) = 1 - p(C_1|\mathbf{x}) \tag{3}$$

You need to know that the MLE parameters for logistic regression can be found by gradient descent.

5 Neural networks

You need to know what the following terms mean:

- Weights
- Activation function
- Input layer
- Hidden layer
- Output layer
- Cost function / Loss function
- Forward pass
- Backward pass
- Backpropagation
- Vanishing gradient problem
- Exploding gradient problem
- Gradient clipping
- Non-saturating activation functions
- Residual layer / network
- Parameter initialisation
- Early stopping
- Weight decay
- Dropout

You need to know that a unit j in a neural network (but not in the input layer) computes a value z_j by first computing a_j, a weighted sum of its inputs (from the previous layer), and then sending a_j to some nonlinear *activation function* h:

$$a_j = \sum_i w_{ji} z_i \tag{4}$$

$$z_j = h(a_j) \tag{5}$$

$$\tag{6}$$

You need to know the *backpropagation formula*:

$$\delta_j = h'(a_j) \sum_k w_{kj} \delta_k \tag{7}$$

and be able to explain what each of the symbols in this formula represents.

6 Trees

You need to know...

- what a classification tree is
- what a regression tree is
- how trees partition the input space
- that trees are a nonparametric method
- how the standard CART algorithm for learning trees works, including the final pruning stage

7 Kernels and SVMs

You need to know what the following terms mean

- kernel function
- the kernel trick
- dual parameter
- Gram matrix
- the margin
- support vectors
- a soft margin
- slack variables

You need to know...

- the role of the regularisation parameter in soft margins
- how SVMs can be extended to deal with having more than two classes

8 Probabilistic Graphical Models

You need to know what the following terms mean

- Directed acyclic graph
- Conditional independence
- Bayesian network
- the structure of a Bayesian network

- the parameters of a Bayesian network
- child (in a Bayesian network)
- parent (in a Bayesian network)
- descendant (in a Bayesian network)
- path (in a Bayesian network)
- collider (in a Bayesian network)
- blocked path (in a Bayesian network)

You need to know...

- the factorisation of a joint probability distribution defined by the structure of a given Bayesian network
- how to use plate notation to compactly represent a Bayesian network
- how to translate a machine learning model (described in words) to a Bayesian network
- how to use d-separation to check for conditional independence relations in a Bayesian network.

9 Bayesian machine learning

You need to know what the following terms mean

- Prior distribution
- Likelihood
- Posterior distribution

You need to know that in the Bayesian approach: the parameters, the data and any unobserved (latent) variables are all represented as random variables in a joint probability distribution. Unknown quantities (parameters and latent variables) are unobserved random variables, know quantities (the data) are observed random variables.

10 Sampling and MCMC

You need to know what the following terms mean

- ancestral sampling
- rejection sampling

- Markov chain
- homogeneous Markov chain
- initial distribution (in a Markov chain)
- transition distribution (in a Markov chain)
- Markov chain Monte Carlo
- target probability distribution
- Metropolis-Hastings algorithm
- Metropolis algorithm
- proposal distribution
- acceptance probability
- burn-in
- convergence (in context of MCMC)

You need to know...

- the equations for the acceptance probability for both the Metropolis and Metropolis-Hastings algorithms.
- how a sample from a distribution can be used to approximate an expected value defined by that distribution
- that in MCMC we sample from a **sequence** of distributions and that the samples are not independent
- why we throw away samples in the burn-in
- why we typically run several chains when doing MCMC
- that $\hat{R}$ is a value computed from an MCMC run used to check for convergence; if the run has been successful (i.e. there's been convergence) it will be close to 1.

11 k-means and Gaussian mixtures

You need to know what the following terms mean

- clustering
- soft clustering
- Gaussian mixture model

- mixing coefficient
- responsibility (in context of a mixture model)

You need to know...

- how the k-means algorithm works
- that one can do soft clustering by applying MLE to a Gaussian mixture model

12 The EM algorithm

You need to know that

- the EM algorithm is an iterative algorithm that attempts to find a value of $\boldsymbol{\theta}$ that maximises the *log-likelihood*: $\ln p(\mathbf{X}|\boldsymbol{\theta})$, where $\mathbf{X}$ is observed data.
- there is no guarantee that the EM algorithm will succeed in maximising the log-likelihood. It may converge to a local maximum which is not a global maximum of the log-likelihood function.

If you are given any or all of the following three EM-related equations:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p)$$

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\ \mathrm{KL}(q||p) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \end{aligned}$$

you should be able to explain what each of the symbols in these equations represent. It can be helpful to you to simply memorise these three equations.

You need to know that

- $\mathrm{KL}(q||p) \geq 0$ for any choice of q, so $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$.
- In the E-step we increase $\mathcal{L}(q, \boldsymbol{\theta})$ by updating q (and leaving $\boldsymbol{\theta}$ fixed).
- In the M-step we increase $\mathcal{L}(q, \boldsymbol{\theta})$ by updating $\boldsymbol{\theta}$ (and leaving q fixed).
- After the E-step we have $\mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta})$ (and so $\mathrm{KL}(q||p) = 0$), so that in the following M-step increasing $\mathcal{L}(q, \boldsymbol{\theta})$ will also increase $\ln p(\mathbf{X}|\boldsymbol{\theta})$.