ML1 - Exam Prep 2022

Rob Hesselink

October 2022

1 General Skills

There are several core skills that you should be able to do for almost any model we discussed. These include

- Vector calculus/componentwise differentiation.
- probability theory, such as marginalisation, pulling joints into conditionals etc.
- Write down the likelihood function, using proper assumptions like iid.
- Find Maximum-likelihood/Maximum-a-posteriori point estimates.
- Constrained optimisation—we'll get into this later.

2 General Knowledge

For every model you should know a couple of things:

- What is it for? Regression or classification?
- How do I train it? Closed-form solution, SGD, EM-algorithm? Can I say something about the complexity of training the model—$O(D^3)$ for linear basis function models.
- Is it linear or non-linear? And in its parameters or its input? Which datasets can/can it not separate? Also, think about naive-bayes and marginals.
- Can this model reason about the uncertainty of its prediction? Bayesian models, naive Bayes, GPs can, most others can't.

This is especially useful for the multiple-choice questions. They tend to test this kind of knowledge.

3 How To: Probabilistic Models

Examples: Any explicit distribution (ML or MAP), naive-Bayes, linear (basis function) regression.

Almost all models we covered are probabilistic models and involve maximisation of some likelihood. The steps here are always the same:

1. Write down the likelihood (with prior if applicable)
2. Change it into the log-likelihood
3. Take the derivative w.r.t. some parameter and set to zero. Don't forget the constraints if there are any
4. Solve for the relevant variable to get θ_{ML} or θ_{MAP}

As a general strategy, you are aiming to increase the number of products in the likelihood as much as you can. We saw this a couple of times in the course, for example:

- $\prod_n$ for iid assumption
- $\prod_d$ for naive Bayes
- $\prod_k ()^{t_{nk}}$ for classification problems, where t_{nk} is a one-hot vector

4 How To: Mixture Models

Mixture models are the strange exception, since they are always of the form $\prod_n \sum_k p(x_n|k)$. When taking the log, you will get the log of the sum, which is nasty and requires the EM algorithm. Always try to find the responsibilities $p(k|x_n)$—that is the E-step—and then maximise over the parameter, yielding an answer in terms of these responsibilities—the M-step.

For any algorithm, you initialise either responsibilities or parameters and then iterate. Always converges to a local optimum.

5 How To: PCA

: PCA questions generally boil down to the same thing: prove properties about the mean and covariance of the projected vectors. Show that they have zero mean and identity covariance, for example. remember that:

- E[X + c] = E[X] + c
- Cov(X + c, Y) = Cov(X, Y)
- Zero-mean data stays zero-mean
- Eigenvectors are orthogonal/orthonormal

- The variance of the projections are the corresponding eigenvalues, so if you want identity covariance (whitening) use $U\Lambda^{-1/2}$

6 How To: SVMs

Two things change the sign in inequality constraints: max/min and $\geq / \leq$. Be kind to yourself and always write inequality constraints as $g \geq 0$. Then you only need to worry about min/max. Where MINimising yields a MINus sign.

KKT's come in threes! There's always the original condition, the positivity for the associated multiplier and then the combo condition. Your answer should always be divisible by three and generally has the shape $3N$.

The goal is to go from the primal Lagrangian to the dual Lagrangian. This is because of two reasons

1. We can use the kernel trick in dual space
2. The dual optimisation problem is guaranteed convex

Primal variables are things you can point to + weights. They have an immediate physical interpretation, such as a radius, or a centroid. Dual variables are generally Lagrange multipliers.

This question also always follows the same shape

1. Write down the Lagrangian, introducing multipliers
2. Write down the KKT conditions
3. Get the stationary conditions by deriving to the primal variables.
4. Use the stationary conditions to eliminate primal variables and create the dual Lagrangian.
5. Given solutions for dual variables, explain how the possible values for the duals relate to regions in the original problem. e.g. slack ¿ 1 means misclassified.