ML1 Tips and Tricks

Sharvaree Vadgama & Rob Hesselink

October 2021

1 MC Questions

There are several types of questions that tend to pop up in every exam, but in different shapes and forms. These are a couple of trends I have noted:

- Which model is suited to this task? Tests whether you understand which models are/can be used for regression, which for classification.
- Which models are (un)supervised?
- Which model can separate this dataset? This generally tests your knowledge of which models have linear decision boundaries and which have non-linear decision boundaries.
- When training model X, are we guaranteed to find a global optimum? This question tests whether you understand how a model is trained. Is there stochasticity involved? And if so, does that mean we reach a different solution every time? Is there a closed-form solution available to us, like in linear regression? And by the way, is that still true for logistic regression?
- Have you developed good intuition for overfitting/bias-variance tradeoff and can you tell me which of these models is overfitting?
- Which kernel function is suitable for this problem? How does it change as we vary the parameters?

2 Vector Calculus

- Three-step plan: (i) Component notation (ii) Differentiate scalarly (iii) infer shapes.
- The number of indices is conserved, except when summmed over. Also, the number of indices tells you what kind of object your derivative is.
- Always differentiate with respect to different indices and let Kronecker delta's do the work.

3 Likelihoods

In general, you will be able to factorise a likelihood. This makes optimisation much easier, since many terms will disappear when taking the log and the derivative. Some common ones include:

- $p(X) = \prod_N p(x_n)$ because i.i.d. assumption.
- $p(x) = \prod_D p(x_d)$ because of naive Bayes assumption.
- $p(x,t) = p(t)p(x|t) = \prod_K [p(t=k)p(x|t=k)]^{t_k}$. This is just a trick useful for multi-class classification.

4 Latent Variable Models

If there is either a sum or an integral in your likelihood, some variable is marginalised over. This means you're dealing with a latent variable model. These come in several forms, one of which is the mixture model.

If you ever see one of these in your likelihood, you should have a knee-jerk reaction: You should compute the posterior over the latent variable $p(z|x)$, which are called the responsibilities. You should then use this in the EM algorithm to iteratively optimise your parameters.

Also, if one of the parameters you're optimising over is a probability itself, don't forget your Lagrange multipliers.

5 SVMs

Solving constrained optimization ($\max_x f(x)$) problem via method of Lagrange multipliers for equality constraints ($g(x) = c$): It can be considered a three step process:

1. Identify the constraint. (Given in the question explicitly or need to be identified from the problem example: it is a probability and hence sums to 1)
2. Define Lagrangian $\mathcal{L}$:

   $$\mathcal{L} = f(x) + \mu_1(constraint1) + \mu_2(constraint2)\dots$$

   Note: If there are n constraints, then we will need n Lagrangian multipliers.
3. Find stationary points of $\mathcal{L}(x, \mu_1, \mu_2 ...)$ Take derivative w.r.t x, set it to zero and find solution for x.

Constrained optimization with inequality constraint: This can be a thought of a six step process:

1. Identify the constraints. (Note: suppose we have two inequality constraints, $g(x) \geq 1$ and $h(x) \leq 0$, remember to swap signs when adding in the $\mathcal{L}$ term.)
2. Define Primal Lagrangian $\mathcal{L}$

$$\mathcal{L} = f(x) + \mu_1(g(x) - 1) - \mu_2 h(x)$$

   where μ_1 and μ_2 are Lagrange multipliers for the constraints.
3. Identify KKT conditions;
   - $\mu_1 \geq 0, \mu_2 \geq 0$
   - $g(x) - 1 \geq 0, -h(x) \geq 0$
   - $\mu_1(g(x) - 1) = 0, -\mu_2 h(x) = 0$
4. Compute dual Lagrangian $\tilde{\mathcal{L}}(\mu)$, optimize w.r.t primal variable x for a fixed dual variable μ.
5. Solve dual problem

$$\mu^* = \underset{\mu}{argmin}\ \tilde{\mathcal{L}}(\mu) \quad \text{subject to} \quad \mu \geq 0$$

6. Maximize primal Lagrangian

$$x^* = \underset{x}{argmax}\ \mathcal{L}(x, \mu^*)$$

Note: If the problem changes from maximization to minimization the signs change. If the inequality is of the form $h(x) \leq 0$ the signs change as given above.

6 Principal Component Analysis

There are several things that are important. For example, if the covariance is expressed as

$$\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,$$

then the columns of $\mathbf{U}$, namely $\mathbf{u}_k$, are orthogonal. Also remember that you're trying to maximise the variance left in the dataset. It's good to remind yourself of an ellipsoid, so that axes align with the variance.

Important questions from previous exams:

- 2020 exam: Q27 (Kernel Ridge Regression)
- 2018 exam: Q2, Q3, Q5 (Maximum margin classifier- square and diamond shaped decision boundaries)
- 2017 exam: Q2, Q4f

Important questions from Practicals:

- P1] 1b,h 2,3
- P2] 1,2
- P3] 1,2
- P4] 1,2
- P5] 2

Important assignment questions:

- A1] 1a,d,e 2d 3c
- A2] 2 4c,d
- A3] 1
- A4] 1 2
- A5] 2