PGMs week 8 Learning: Maximum Posterior Estimation

- Watch the 16.3 Koller video.
- Re-read Barber chapter 8, as well as chapter 9. Skip section 9.5 (structural learning).
- Important points:
  - The model parameters now become random variables in own right, described by a *prior* probability distribution $p(\boldsymbol{\Theta})$.
  - MAP estimation sets the model parameters $\boldsymbol{\Theta}$ in such a way that the posterior probability of the parameters given the data is maximised, i.e. find

$$\begin{aligned}\boldsymbol{\Theta}^{\text{MAP}} &= \overset{\text{argmax}}{\boldsymbol{\Theta}} \quad p(\boldsymbol{\Theta}|\mathcal{D}) \\ &\simeq \overset{\text{argmax}}{\boldsymbol{\Theta}} \quad p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}). \qquad (1)\end{aligned}$$

  - Being a distribution, the prior distribution in its turn is governed by its own parameters, now known as *hyper-parameters*. Closer inspection should show that these hyper-parameters are exactly what we tune when we set regularisation parameters. But now they have physical meaning in terms of the constraints they place on the model parameters.
  - These hyper-parameters are normally tuned using a validation set, but ...
  - nothing prevents us from considering them to be random variables in their turn too, and instead tune their hyper-hyper-parameters on a validation set. Or at this level of abstraction, simply choose them to be something sensible.
  - It is often mathematically convenient to choose the form of the prior to match that of the posterior - this makes of it simply another term similar to the others in the likelihood product. This is called a conjugate prior. For multinomials (i.e. probability tables) we use the Dirichlet prior. The joint prior for the mean and precision of a 1-dim Gaussian is the Gaussian-gamma distribution, and for a multi-dim Gaussian it is the Gaussian-Wishart distribution.
  - Using the inference techniques we already encountered, we in principle can infer the probability distribution of our model parameters. In the full Bayesian approach we know the parameters as a *distribution* instead of a set of values. When observing discrete data this is in principle always possible, although computational cost may preclude it.
  - With continuous features, the required integration required for marginalisation often makes this impossible to do directly. One of our alternatives is to switch to max-sum inference instead of sum-product inference. The maximisation can then be done by finding the maximal points on the respective distributions – this can be done via differentiation/optimisation, an easier alternative to the integrations originally required. In effect this finds the most likely parameter values for the model given the observed data.
  - The resultant MAP estimation equations often take the form where it appears as if the training data is supplemented with a number of extra “ghost” features which correspond to the characteristics of the prior.
- Exercise:
  1. Repeat the Gaussian estimate of last week, but this time as a MAP estimate. Specifically write the estimation equations in the form where it appears as if there are extra training feature vectors. Note how, with zero real training vectors the prior fully determines the resultant estimate (and therefore also stabilises it), while with infinitely many training vectors the MAP estimate reduces to an ML estimate.
  2. Repeat the logistic regression task from last week, now using a zero-mean Gaussian with uniform precision λ as the prior for the regression weights.
     (a) Confirm that this reduces to exactly the same regularisation form that is normally used for this model.

(b) If we were to use a hyper-prior for λ, what would its distribution be (i.e. consult Barber and other sources to find out what the conjugate prior for a Gaussian precision should be). What does the corresponding regularisation term now look like?

(c) Train and test the logistic regression classifier using both of these approaches. Compare.