

## CS4780/5780 Homework 3

Due: Tue. 10/02/2018 11:55pm on Gradescope

### Problem: Estimating Probabilities from data

Suppose we are given a dataset  $\mathcal{D} = \{\vec{x}_1, \dots, \vec{x}_n\} \subset \mathbb{R}^d$ . We would like to model the data with a Gaussian distribution with an unknown  $\vec{\mu}$  and a known covariance matrix,  $I$ , namely

$$\mathcal{P}(\vec{x}; \vec{\mu}, I) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^T (\vec{x} - \vec{\mu})\right)$$

For estimation purposes, suppose each data point  $\vec{x}_i$  is drawn independently from the above Gaussian distribution.

1. Show how we could use Maximum Likelihood Estimation (MLE) to estimate  $\vec{\mu}$  from the data  $\mathcal{D}$  by maximizing the log likelihood function,  $\log P(\mathcal{D}|\vec{\mu})$ . Recall that in order to find the maximizer of a function you need to set its gradient to zero. Please also compute the second derivative and verify that the point you find is indeed a maximum.
2. One of the problems with MLE is that we cannot introduce any prior belief on  $\vec{\mu}$ . For instance you could not incorporate prior knowledge that  $\vec{\mu}$  should be close to the origin. To work around this, we will use a Bayesian approach, i.e., Maximum A Posterior estimation (MAP). With MAP we can introduce a prior on  $\vec{\mu}$ , defined by a standard Gaussian ( $\vec{\mu} \sim \mathcal{N}(0, I)$ ), to encourage  $\vec{\mu}$  to be close to the origin. We can then use Bayes Rule to solve for the most likely  $\vec{\mu}$  given the data and our prior belief:

$$\vec{\mu} = \arg \max_{\vec{\mu}} P(\vec{\mu}|\mathcal{D}) = \arg \max_{\vec{\mu}} \frac{P(\mathcal{D}|\vec{\mu})P(\vec{\mu})}{P(\mathcal{D})}. \quad (1)$$

Find the solution for (1), following steps similar to those shown in class. (Hint: You will not need to compute  $P(\mathcal{D})$ .)

3. MAP is cool, but if you want to be really fancy you may be interested in computing the true posterior distribution,  $P(\vec{\mu}|\mathcal{D})$ , and not just its maximum. This can often be difficult, but for Gaussian distributions you can compute the posterior in closed form. Assume that we are going to place the same prior on  $\vec{\mu}$  as in part 2. We know that by Bayes' rule:

$$P(\vec{\mu}|\mathcal{D}) \propto P(\mathcal{D}|\vec{\mu})P(\vec{\mu})$$

Since the product of two Gaussian distributions is a Gaussian distribution, the posterior has to be a Gaussian distribution too, i.e.  $\vec{\mu} \sim \mathcal{N}(\vec{m}, \Sigma)$ . Using this fact we can assume that

$$P(\vec{\mu}|\mathcal{D}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\vec{\mu} - \vec{m})^T \Sigma^{-1} (\vec{\mu} - \vec{m})\right) \quad (2)$$

and we can simplify the procedure of finding the posterior to finding  $\vec{m}$  and  $\Sigma$ . Find the parameters  $\vec{m}$  and  $\Sigma$  by comparing the exponents in  $\mathcal{N}(\vec{m}, \Sigma)$  and  $P(\mathcal{D}|\vec{\mu})P(\vec{\mu})$ .