Deep|Bayes summer school 2019: homework

Bayesian methods problems

1. (Bayesian reasoning) During medical checkup, one of the tests indicates a serious disease. The test has high accuracy 99% (probability of true positive is 99%, probability of true negative is 99%). However, the disease is quite rare, and only one person of 10000 is affected. Calculate the probability that the examined person has the disease.

2. (Data modeling: Bayesian vs frequentist) Let $X = \{x_1, \ldots, x_N\}$ be N independent coin tosses, $x_i \in \{0, 1\},\ i = 1, \ldots, N$. If $\theta \in [0, 1]$ denotes the probability of landing heads up, the likelihood (Bernoulli distribution) has the form

$$p(X \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta), \quad p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i},\ i = 1, \ldots, N.$$

   We would like to estimate the parameter θ in a frequentist and Bayesian way. In order to perform an analytical Bayesian inference, we choose a conjugate prior. The conjugate prior distribution for the Bernoulli likelihood is a Beta distribution:

$$p(\theta \mid a, b) = Beta(\theta \mid a, b) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}, \quad a > 0,\ b > 0$$

   where $B(a, b)$ denotes Beta-function (normalizing constant).

   (a) Compute the maximum likelihood estimate for θ.
   (b) Check that the Beta distribution is indeed the conjugate distribution for the Bernoulli likelihood.
   (c) Compute the posterior distribution $p(\theta \mid X, a, b)$.
   (d) Compute the expectation of the posterior distribution and compare it with the maximum likelihood estimate.
   (e) Compute the posterior predictive distribution $p(x_{N+1} = 1 \mid X, a, b) = \int_{[0,1]} p(x_{N+1} = 1 \mid \theta)p(\theta \mid X, a, b)\mathrm{d}\theta$.
   (f) Which a and b will you choose if you think that the coin is fair? If you think the coin is unfair and has heads on both sides?

3. (Bayesian inference: analytical vs approximate) Consider someone making everyday notes $X = \{x_1, \ldots, x_N\}$ on how many seconds the train is late or early for. Let's assume that

$$p(X \mid \mu,\, \lambda) = \prod_{i=1}^{N} p(x_i \mid \mu,\, \lambda), \quad p(x_i \mid \mu,\, \lambda) = \mathcal{N}(x_i \mid \mu,\, \lambda^{-1}),\ i = 1, \ldots, N.$$

   We would like to perform a Bayesian inference on parameters μ and λ, i.e. find the posterior distribution $p(\mu,\, \lambda \mid X)$. Let's consider different priors and in each case use an appropriate inference type.

   (a) If we choose a conjugate prior then we can perform analytical Bayesian inference. A conjugate prior to the normal likelihood is a normal-gamma distribution (let's choose particular prior parameters for brevity):

$$p(\mu,\, \lambda) = \mathcal{NG}(\mu, \lambda \mid 0, 1, 1, 1) = \mathcal{N}(\mu \mid 0,\, \lambda^{-1})\mathcal{G}(\lambda \mid 1, 1).$$

   Check that $p(X \mid \mu,\, \lambda)$ and $p(\mu,\, \lambda)$ are conjugate and find the posterior distribution $p(\mu,\, \lambda \mid X)$.

   (b) If we choose not a conjugate but a conditionally conjugate prior (broader class):

$$p(\mu,\, \lambda) = p(\mu)p(\lambda) = \mathcal{N}(\mu \mid 0,\, 1)\mathcal{G}(\lambda \mid 1, 1),$$

   we can perform variational inference with mean-field approximation:

$$p(\mu,\, \lambda \mid X) \approx q(\mu)q(\lambda).$$

   Check that the prior is conditionally conjugate to the likelihood and find $q(\mu)$ and $q(\lambda)$.

(c) If we decide to perform a Bayesian inference only for μ, and for λ use a maximum likelihood estimate, then we can choose a conjugate prior on μ:

$$p(\mu) = \mathcal{N}(\mu \mid 0, 1)$$

and use EM-algorithm. Check that $p(X \mid \mu, \lambda)$ and $p(\mu)$ are conjugate when λ is fixed. Derive formulas for EM-algorithm: compute the posterior $p(\mu \mid X, \lambda)$ assuming λ is fixed (E-step) and find optimal λ assuming $p(\mu \mid X, \lambda)$ is fixed (M-step).

In this task we use the following parametrizations for normal and gamma distributions:

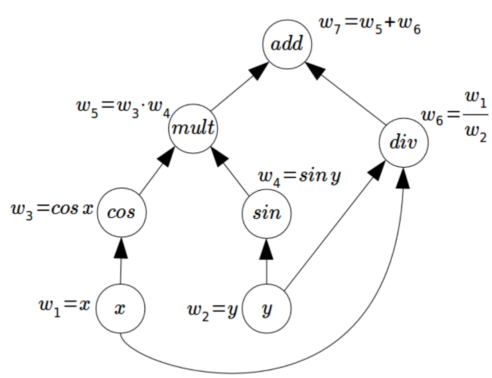
$$\mathcal{N}(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2}(x-\mu)^2}, \quad \mathcal{G}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda).$$

Deep learning problems

1. Write formulas for:
   - 1 step of stochastic gradient descend with mini-batch size of 1;
   - forward pass through fully-connected, convolutional and recurrent layer;
   - 2-dimensional image convolution;
   - forward pass through a dropout layer for training and testing stage;
   - forward pass through a batch normalization layer;
   - loss for the generative adversarial network.

   Make sure you understand all the notation.

2. Find the derivatives of w_7 with respect to x and y using backpropagation algorithm in the following computational graph:



3. Propagate gradients through the linear layer $y = Wx$: given $\frac{\partial L}{\partial y}$, find $\frac{\partial L}{\partial x}$ and $\frac{\partial L}{\partial W}$. Dimensions: $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $W \in \mathbb{R}^{m \times n}$.

4. The size of the input image is $21 \times 11 \times 3$ (height $\times$ width $\times$ number of channels). You apply a 2-dimensional convolution with kernel size 5×3, no padding and stride 2. What is the size of the output image?

5. What are the specific features of VGG architecture? Of ResNet?

6. What is the motivation to use batch normalization? How does it help training?

7. What are the problems of using generative adversarial networks?