Midterm Exam

Q1: If two binary random variables X and Y are independent, are $\bar{X}$ (the complement of X) and Y also independent? Prove your claim.

A: If $\Pr[XY] = \Pr[X]\Pr[Y]$, then

$$\Pr[\bar{X}Y] = \Pr[Y - XY] = \Pr[Y] - \Pr[X]\Pr[Y] = \Pr[\bar{X}]\Pr[Y].$$

So $\bar{X}$ and Y are independent.

Q2: To estimate the head probability θ of a coin from the results of N flips, we use fictional observations (or pseudo-counts) to incorporate our belief of the fairness of the coin. This is equivalent to using which distribution as a prior of θ?

A: The coin flipping satisfies Bernoulli distribution, so if the prior can be incorporated into it, the prior is equivalent to a Bernoulli distribution.

Q3: Suppose we have two sensors with known (and different) variances v_1 and v_2, but unknown (and the same) mean μ. Suppose we observe N_1 observations $y_i^{(1)} \sim N(\mu, v_1)$ from the first sensor and N_2 observations $y_i^{(2)} \sim N(\mu, v_2)$ from the second sensor. Let D represent all the data from both sensors. What is the posterior $p(\mu|D)$ assuming a noninformative prior for μ (which we can simulate using a Gaussian with a precision of 0)?

A: The noninformative prior is a constant $p(\mu) = \mu$.

The likelihood of the first sensor is

$$p(D_1|\mu) \propto \exp(\frac{N_1(\bar{D}_1 - \mu)^2}{2v_1^2}).$$

Because of the constant prior, the posterior is also proportional to it.

The posterior of the second sensor is the same way, which is

$$p(\mu|D_2) \propto \exp(\frac{N_2(\bar{D}_2 - \mu)^2}{2v_2^2}).$$

So the posterior on D is

$$p(\mu|D) \propto \exp(\frac{N_1(\bar{D}_1 - \mu)^2}{2v_1^2} + \frac{N_2(\bar{D}_2 - \mu)^2}{2v_2^2}).$$

This should be a Normal distribution, from which the μ can be solved.

Q4: About undirect graphical model.

A: (a) The Markov properties of the random variables are

$$\begin{aligned} p(x_1|x_2, x_3, x_4) = p(x_1|x_2, x_4), \quad p(x_2|x_1, x_3, x_4) = p(x_2|x_1, x_3) \\ p(x_3|x_1, x_2, x_4) = p(x_3|x_2, x_4), \quad p(x_4|x_1, x_2, x_3) = p(x_4|x_1, x_3). \end{aligned}$$

(b) The cliques are $\{x_1, x_4\}, \ \{x_1, x_2\}, \{x_2, x_3\}, \{x_3, x_4\}$. So the factorization of the joint distribution $p(x_1, x_2, x_3, x_4)$ w.r.t. $\psi(\cdot)$ is

$$p(x_1, x_2, x_3, x_4) = \psi(x_1, x_4)\psi(x_1, x_2)\psi(x_2, x_3)\psi(x_3, x_4).$$

Q5: ML estimation for linear regression.

A: If x_i, y_i are regraded as the instantiations of random variables X, Y, then

$$\lim_{N\to\infty} \bar{x} = \mathrm{E}[X], \quad \lim_{N\to\infty} \bar{y} = \mathrm{E}[Y], \quad \lim_{N\to\infty} \sum_i^N \frac{x_i y_i}{N} = \mathrm{E}[XY].$$

When N is large, it has

$$\frac{\mathrm{cov}[X, Y]}{\mathrm{var}[X]} = \frac{\mathrm{E}[XY] - \mathrm{E}[Y]\mathrm{E}[Y]}{\mathrm{E}[X^2] - \mathrm{E}[X]^2} \approx \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2},$$

and

$$\bar{y} - w_1\bar{x} \approx \mathrm{E}[Y] - w_1\mathrm{E}(X).$$

Q6: The logit of the probability of a logistic regression model is a linear function of x.

A: The nominator is

$$p(y=1|x,w) = sigm(w^T x + b) = \frac{1}{1+\exp(-w^T x - b)}.$$

The denominator is

$$p(y=0|x,w) = 1 - sigm(w^T x + b) = \frac{\exp(-w^T x - b)}{1+\exp(-w^T x - b)}.$$

So the logit of them

$$\ln\left(\frac{p(y=1|x,w)}{p(y=0|x,w)}\right) = \ln\left(\frac{1}{\exp(-w^T x - b)}\right) = w^T x + b$$

is a linear function of x.

Q7: Gaussian mixture model vs k-means algorithm.

A: (a) True. Uniform $p(z) = 1/k$ is an assumption of the k-means algorithm, which means that the probability of each sample coming from any cluster is the same; in GMM this can be any distribution.

(b) True. k-means assumes that features are independent of each other. As a result, the covariance between different features tends to zero.

(c) True: The covariance matrix in k-means is assumed as $\Sigma = I$.

Q8: Optimal log-likelihood of Bernoulli.

A: (a) The objective function is

$$l(\theta) = -\ln\left(\prod f(x_i|\theta)\right) = -\ln\left((1-\theta)^2\theta\right) = -(2\ln(1-\theta) + \ln\theta).$$

When $l'(\theta) = 0$, i.e.

$$l'(\theta) = -\left(\frac{-2}{1-\theta} + \frac{1}{\theta}\right) = 0 \quad \rightarrow \quad \theta = \frac{1}{3}.$$

This is the global minima of $l(\theta)$, under the constraint $\theta \geq 1/2$, the optimal solution is $\theta = 1/2$.

(b) The inequality constraint can be written as $\theta - 1/2 - s^2 = 0$, where $s \geq 0$ is a slack variable. Then the Lagrangian function is

$$L(\theta, \lambda, s) = l(\theta) - \lambda(\theta - \frac{1}{2} - s^2),$$

where $\lambda \geq 0$ is the multiplier.