Assignment 3: Generalized Linear Regression and Graphical Models

---

Q1: MAP estimation for 1D Gaussian.

A: (a) The posterior of the unknown mean is

$$p(\mu|x) \propto p(x|\mu) \times p(\mu),$$

which is a Gaussian distribution. The MAP of μ is the mode (also the mean), which is given by

$$\hat{\mu} = (\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}) \times \sigma_N^2, \quad \text{with } \sigma_N^2 = \frac{\sigma^2\sigma_0^2}{\sigma_2 + N\sigma_0^2}.$$

(b) The MLE of μ is $\mu = \bar{x}$. As the increase of N in the above formula, item $N\bar{x}$ dominates the result, so it converges to $\bar{x}$

(c) As the increase of σ_0^2, the item μ_0/σ_0^2 converges to zero, so the MAP also converges to the MLE.

(d) As the decrease of σ_0^2, the result is dominated by μ_0 item, so the MAP converges to the prior.

Q2: Optimizer of $l(w)$ with regularization.

A: The result can be proved by calculating the derivative of $l(s)$

$$\frac{\mathrm{d}l}{\mathrm{d}w} = 2X^T(Xw - y) + 2\lambda w.$$

When the derivative equals zero, it can be solved that $w = (X^TX + \lambda I)^{-1}X^Ty$.

Q3: About logistic regression.

A: (a) False. The form of $l(w, D)$ is

$$l(w, D) = \frac{1}{N}\sum_{i=1}^{N} -log(1 + \exp(-y_i x_i^T w)).$$

It is a convex function w.r.t w, so there is a global optimal.

(2) False. L_2-norm regularization is a smooth function that does not tend to give sparse solutions.

(3) False. $l(w, D)$ is the log-likelihood, so as the increase of regularization, the log-likelihood becomes smaller.

(4) False. The same reason as the above.

Q4: One-dimensional linear regression.

A: (a) The log-likelihood of w, σ^2 is

$$l(w, \sigma^2) = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}RSS(w),$$

where $RSS(w)$ is the relative square sum $(y - Xw)^T(y - Xw)$. From this, the MLE estimate of w can be calculated as

$$\hat{w} = (X^TX)^{-1}X^Ty.$$

Then $\hat{w}$ is calculated as 0.0126 and $\hat{\sigma}^2$ as 0.1513.

(b) When w has a prior $p(w) = \mathcal{N}(w|0, 1)$, the posterior is also a Gaussian distribution with

$$p(w|X, y, \hat{\sigma}^2, 0, 1) \propto \mathcal{N}(w|w_N, \Sigma_N),$$

where the posterior mean w_N is

$$w_N = (X^TX + 1)^{-1}(X^TX\hat{w} + 0) = 0.0126$$

from the Bayesian update rule for Gaussian conjugate prior. (This result needs to be checked. I am not sure whether it is correct or not.)

Q5: Properties of the sigmoid function.

A: (a) The equation can be proved by

$$\frac{\mathrm{d}\mu}{\mathrm{d}a} = \frac{\exp(-a)}{(1 + \exp(-a))^2} = \frac{1}{1 + \exp(-a)} \cdot \frac{\exp(-a)}{1 + \exp(-a)}.$$

(b) The negative log-likelihood is

$$l(w) = -\sum_{i=1}^{N} y_i \log \mu(x_i) + (1 - y_i)\log(1 - \mu(x_i)).$$

Then, its gradient is

$$
\begin{aligned}
g(w) = \frac{\mathrm{d}l(w)}{\mathrm{d}w} &= -\sum_{i=1}^{N}\left(\frac{y_i}{\mu(x_i)}\frac{\mathrm{d}\mu(x_i)}{\mathrm{d}w} - \frac{1-y_i}{1-\mu(x_i)}\frac{\mathrm{d}\mu(x_i)}{\mathrm{d}w}\right) \\
&= \sum_{i=1}^{N} x_i(\mu(x_i) - y_i) = X^T(\mu - y).
\end{aligned}
$$

(c) Because $\mu(1-\mu) \in (0,1)$, so S is positive-definite. $X^T X$ is semi-positive-definite, then H is semi-positive-definite.

Q6: About Bayesian network.

A: (a) Equivalent. For example, $A \perp C \mid B$ holds for the two BNs.

(b) Not equivalent. For example, $A \perp C \mid B$ in the second BN, but this is not true in the first.

(c) Equivalent. B, C, D are the same for them; besides this, A only depends on B. There is an independence assumption $A \perp C \mid B$.

(d) Not equivalent. They have different structures. $B \perp D \mid C$ holds in the first graph, but it does not hold in the second.