Hw1_writeup

Sunday, October 9, 2016 8:02 PM

1c

It is almost same as gradient descent but only takes small batches to estimate using a batch from the input to estimate the cost function.

making the random batches from the training set.

Pseudo code:

```
for i = 0 ... N do:
    initialize d_i randomly
    let t = 1
    for x in training_batches:
        if (t ≤ T and stopping condition not True):
            A_i ← x^T x − Σ_{j=0}^{i-1} λ_j d_j d_j^T and
            y ← d_i − η ∇_{d_i}(−d_i^T A_i d_i)
            d_i ← y / ||y||
            t ← t+1
        else: return
    λ_i ← d_i^T x^T x d_i
```

Here A_i is estimated using the batches of training set rather than the whole training set.

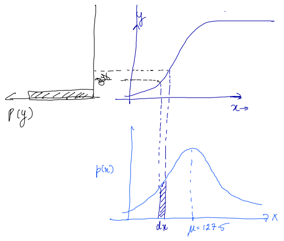
THEORETICAL:-

2.d (i)

$$p(x) = \frac{f(x)}{\int_0^{255} f(x)\,dx} \approx f(x)$$

where $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$\sigma^2 = 1$, $\mu = 127.5$



For any given mapping $y = g(x)$ between x and y we would have:

$$\boxed{p(y)dy = p(x)dx}$$

where - $0 \le x \le 255$

$$\int_0^{255} p(x)dx = 1$$

$$p(y) = \frac{1}{255}$$

and

$$y = g(x)$$

$$dy = d(g(x))$$

$$\Rightarrow d(g(x))_x = 255\,p(x)\,dx$$

$$\Rightarrow \int_0^x d(g(x)) = 255 \int_0^x p(x)\,dx$$

$$\Rightarrow g(x) = 255F(x)$$ where $F(x)$ is cummulative dist^n f^n of x.

So $\boxed{y = g(x) = 255F(x)}$

(ii)

$$P(X=x, Y=y, Z=z) = \begin{cases} 8xyz & \text{for } x,y,z \in [0,1] \\ 0 & \text{o/w} \end{cases}$$

$$P(X=x) = \int_0^1\int_0^1 p(X=x, Y=y, Z=z)\,dy\,dz$$

$$= \int_0^1\int_0^1 8xyz\,dy\,dz$$

$$= \int_0^1 \left[\frac{8xyz^2}{2}\right]_0^1 dy$$

$$= \int_0^1 4xy\,dy$$

$$= \left[\frac{4xy^2}{2}\right]_0^1$$

$$= 2x$$

lly. $p(Y=y) = \int_0^1\int_0^1 P(X=x, Y=y, Z=z)\,dx\,dz$

$$= \int_0^1 4xy\,dx$$
$$= 2y$$

and $P(Z=z) = \int_0^1\int_0^1 P(X=x, Y=y, Z=z)\,dy\,dx$
$$= \int_0^1 4zy\,dy$$
$$= 2z$$

$T = XYZ$

$T \in [0,1]$

$E(T) = \int_0^1 t\,P(T=t)\,dt$

$$P(T=t) = \int_0^1\int_0^1 P(X=x, Y=y, Z=\tfrac{t}{xy})\,dx\,dy$$
$$= \int_0^1\int_0^1 8xy\left(\tfrac{t}{xy}\right)dx\,dy$$
$$= 8t$$

$$\Rightarrow \quad E[T] = \int_0^1 t(8t)\,dt$$
$$= \left[\frac{8t^3}{3}\right]_0^1$$
$$= \frac{8}{3}$$

$$P(X=x, Y=y \mid Z=z_0) = \frac{P(X=x, Y=y, Z=z_0)}{P(Z=z_0)}$$ (applying Bayes rule).

$$= \frac{\overset{4}{\cancel{8}}xyz_0}{\cancel{2}z_0} = 4xy$$

$$P(X=x \mid Z=z_0) = \frac{\int_0^1 P(X=x, Y=y, Z=z_0)\,dy}{2z_0}$$
$$= \frac{8xz_0}{2z_0}\int_0^1 y\,dy$$
$$= 2x$$

llly $P(Y=y \mid Z=z_0) = \frac{\int_0^1 P(X=x, Y=y, Z=z_0)\,dx}{2z_0}$
$$= 2y$$

$$\Rightarrow \quad P(X=x, Y=y \mid Z=z_0) = P(X=x \mid Z=z_0)\cdot P(Y=y \mid Z=z_0)$$

$\therefore$ They are Conditionally independent.

e (1) $x \sim N(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

$p(\mu \mid \mu_0, \Sigma) = 1 \quad \frac{1}{\;} \quad \exp(-\frac{1}{\;}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0))$

$$p(\mu \mid \mu_0, \Sigma_0) = \frac{1}{(2\pi)^{n/2} |\Sigma_0|^{1/2}} \exp\left(-\frac{1}{2}(\mu-\mu_0)^T \Sigma_0^{-1}(\mu-\mu_0)\right)$$

$$\hat{\mu}_{MAP} = \arg\max_{\mu} f(\mu \mid x, \Sigma)$$

where $f(\mu \mid x, \Sigma)$ is posterior distribution p.d.f.

where $x = \{x^{(i)}\}$ $i = 1(1)m$.

$$f(\mu \mid x, \Sigma) = \frac{f(\mu, x \mid \Sigma)}{f(x \mid \Sigma)}$$

$$= \frac{f(x \mid \Sigma, \mu)\, p(\mu \mid \Sigma)}{f(x \mid \Sigma)}$$

$$= \frac{K \cdot \exp\left(-\frac{1}{2}\left[\sum_{i=1}^{m} (\mu - x^{(i)})^T \Sigma^{-1}(\mu - x^{(i)}) + (\mu-\mu_0)^T \Sigma_0^{-1}(\mu-\mu_0)\right]\right)}{\int_{\mu} K \exp\left(-\frac{1}{2}\left[\sum_i (\mu - x^{(i)})^T \Sigma^{-1}(\mu - x^{(i)}) + (\mu-\mu_0)^T \Sigma_0^{-1}(\mu-\mu_0)\right]\right) d\mu} \quad \text{— (1)}$$

where $K = \frac{1}{(2\pi)^{\frac{(m+1)n}{2}}} \frac{1}{|\Sigma_0|^{n/2}} \frac{1}{|\Sigma|^{\frac{mn}{2}}}$

writing expression (1) as (2) with μ_n, Σ_n as f^n of $m, \Sigma, \Sigma_0, \mu_0$

$$= \frac{K'(m, \Sigma_0, \Sigma, x, \mu_0) \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma_n|^{1/2}} \exp\left(-\frac{1}{2}(\mu-\mu_n)^T \Sigma_n^{-1}(\mu-\mu_n)\right)}{\int_{\mu} K'(m, \Sigma_0, \Sigma, x, \mu_0) \underbrace{\frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma_n|^{1/2}} \exp\left(-\frac{1}{2}(\mu-\mu_n)^T \Sigma_n^{-1}(\mu-\mu_n)\right)}_{\text{multivariate normal distribution}} d\mu} \quad \text{— (2)}$$

$$= \frac{\cancel{K'} \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma_n|^{1/2}} \exp\left(-\frac{1}{2}(\mu-\mu_n)^T \Sigma_n^{-1}(\mu-\mu_n)\right)}{\cancel{K'}}$$

$$\Rightarrow f(\mu \mid x, \Sigma) \sim N(\mu_n, \Sigma_n)$$

and the maximum value of $f(\mu \mid x, \Sigma)$ will be at $\mu = \mu_n$.

- μ_n and Σ_n can be found by comparing the coefficients and $\mu^T \Sigma_n \mu$ in following expressions

$$(\mu-\mu_n)^T \Sigma_n^{-1} (\mu-\mu_n), \quad \sum_{i=1}^{m} (\mu - x^{(i)})^T \Sigma^{-1} (\mu - x^{(i)}) + (\mu-\mu_0)^T \Sigma_0^{-1} (\mu-\mu_0)$$

- $$\mu^T \Sigma_n^{-1} \mu = \mu^T (m \Sigma^{-1} + \Sigma_0^{-1}) \mu$$

$$\Sigma_n = (m \Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

comparing coefficients of μ^T

- $$\mu_n = \Sigma_n (m \Sigma^{-1} \bar{x} + \Sigma_0^{-1} \mu_0)$$

$$\hat{\mu}_{MAP} = (m \Sigma^{-1} + \Sigma_0^{-1})^{-1} (m \Sigma^{-1} \bar{x} + \Sigma_0^{-1} \mu_0)$$

11 y. assuming Σ has no prior distribution

$$f(\Sigma \mid x, \mu) = \frac{f(x, \Sigma \mid \mu)}{f(x \mid \mu)}$$

$$\hat{\Sigma}_{MAP} = \underset{\Sigma}{\arg\max} \left(f(\Sigma \mid x, \mu) \right)$$

where f is p.d.f. of Σ.

$$f(\Sigma \mid x, \mu) = \frac{f(x \mid \Sigma, \mu)}{f(x \mid \mu)}$$

$$\Rightarrow f(\Sigma \mid x, \mu) = \frac{\frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \sum_i (x^{(i)} - \mu) \Sigma^{-1} (x^{(i)} - \mu)\right)}{f(x \mid \mu)} \rightarrow \text{not a } f^n \text{ of } \Sigma.$$

we can find the Σ' s.t. $f(\Sigma' \mid x, \mu)$ is max. and $\boxed{\hat{\Sigma}_{MAP} = \Sigma'}$

Since log is an monotonously increasing function. $\therefore$ finding $\underset{\Sigma}{\arg\max} f(\Sigma \mid x, \mu)$ is same as finding $\underset{\Sigma}{\arg\max} \log f(\Sigma \mid x, \mu)$

$\therefore$ MLE and MAP estimator for Σ is same.

and since $f(x \mid \mu)$ not a f^n of Σ

then $\underset{\Sigma}{\arg\max} f(\Sigma \mid x, \mu) = \underset{\Sigma}{\arg\max} (\log f(\Sigma \mid x, \mu))$

$$= \underset{\Sigma}{\arg\max} \left(\log f(x \mid \Sigma, \mu) \right)$$

$\therefore$ MLE and MAP of Σ will be.

$$\hat{\Sigma}_{MAP} = \hat{\Sigma}_{MLE} = \underset{\Sigma}{\arg\max} \left(\log f(x \mid \Sigma, \mu) \right)$$

$$\hat{\Sigma}_{MAP} = \hat{\Sigma}_{MLE} = \underset{\Sigma}{\arg\max}\left(\log f(X \mid \Sigma, \mu)\right)$$

$$f(X \mid \Sigma, \mu)$$

$$= \frac{\prod_{i=1}^{m} \exp\left(-\frac{1}{2}(x^i - \mu)^T \Sigma^{-1}(x^i - \mu)\right)}{(2\pi)^{\frac{mn}{2}} |\Sigma|^{\frac{1}{2}}}$$

$$= (2\pi)^{\frac{mn}{2}} |\Sigma|^{\frac{-m}{2}} \exp\left(-\frac{1}{2}\sum_i (x^i - \mu)\Sigma^{-1}(x^i - \mu)\right) \quad \text{—} (1)$$

now $\quad A = \sum_{i=1}^{m} (x^i - \mu)(x^i - \mu)^T$

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^i$$

as $\underbrace{\sum_{i=1}^{m} (x^i - \mu)^T \Sigma^{-1}(x^i - \mu)}_{\text{Scalar}}$

$$= \sum_{i=1}^{m} (x^i - \bar{x} + \bar{x} - \mu)^T \Sigma^{-1}(x^i - \bar{x} + \bar{x} - \mu)$$

$$= \left[\sum_{i=1}^{m} \underbrace{(x^i - \bar{x})^T \Sigma^{-1}(x^i - \bar{x})}_{\text{Scalar (can take trace)}}\right] + m(\bar{x} - \mu)^T \Sigma^{-1}(\bar{x} - \mu)$$

$$\Rightarrow \operatorname{tr}\left(\Sigma^{-1}\sum_{i=1}^{m} (x^i - \bar{x})(x^i - \bar{x})^T\right) + m(\bar{x} - \mu)^T \Sigma^{-1}(\bar{x} - \mu)$$

$$\left(\operatorname{tr}(AB) = \operatorname{tr}(BA)\right)$$

$\therefore$ ① becomes

$$f(x|\Sigma;\mu) = 2\pi^{-\frac{mn}{2}} |\Sigma^{-1}|^{\frac{m}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}A) - \frac{m}{2}(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu\right.$$

$\hat{\mu}_{MLE} = \bar{x}$

solving for MLE of Σ assuming that μ is $\hat{\mu}_{MLE}$

$\log f(x|\Sigma,\mu))$

$$= -\frac{mn}{2}\log 2\pi + \underbrace{\frac{m}{2}\log|\Sigma^{-1}| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}A)} - \underbrace{\frac{m}{2}(\bar{x}-\mu)^T\Sigma^{-1}(\bar{x}-\mu)}_{=0}$$

$\Downarrow$

$$\underbrace{\frac{m}{2}\log|\Sigma^{-1}A| - \frac{1}{2}\mathrm{tr}\,\Sigma^{-1}A}_{②} - \underbrace{\frac{m}{2}\log|A|}$$

to maximise we can ignore this.

$\lambda_1\ \lambda_2\ ---\ \lambda_k$ be eigen values of $\Sigma^{-1}A$

So: ② $\Rightarrow$

$$\frac{m}{2}\log\left(\prod_{i=1}^{k}\lambda_i\right) - \frac{1}{2}\left(\sum_{k=1}^{k}\lambda_i\right)$$

[This is maximised when each $\lambda_i = m$.]

$\therefore\ \Sigma^{-1}A = P(mI)P^T \rightarrow P = [e_1\ e_2\ ---\ e_k]$ matrix of eigenvectors

$$\Rightarrow \Sigma^{-1} = mA^{-1}$$

$$\Sigma = \frac{1}{m}A$$

$$\Rightarrow \hat{\Sigma}_{MAP} = \hat{\Sigma}_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\bar{x})(x^{i}-\bar{x})^T$$

Monday, October 10, 2016 6:47 PM

$$\hat{\mu}_{MAP} = (m\Sigma^{-1} + \Sigma_0^{-1})^{-1}(m\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0)$$

$$\hat{\Sigma}_{MAP} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T$$

checking whether biased or not

(1) $$\hat{\mu}_{MAP} = (m\Sigma^{-1} + \Sigma_0^{-1})^{-1}(m\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0)$$

$$E(\hat{\mu}_{MAP}) = (m\Sigma^{-1} + \Sigma_0^{-1})^{-1}E(m\Sigma^{-1}\bar{x} + \Sigma_0^{-1}\mu_0)$$

$$= (m\Sigma^{-1} + \Sigma_0^{-1})^{-1}(m\Sigma^{-1}E(\bar{x}) + \Sigma_0^{-1}\mu_0)$$

as $E()$ is linear f^n we can take inside bracket and $\Sigma_0^{-1}\mu_0$ is constant $\therefore E(\Sigma_0^{-1}\mu_0) = \Sigma_0^{-1}\mu_0$.

$$= (m\Sigma^{-1} + \Sigma_0^{-1})^{-1}(m\Sigma^{-1}\mu + \Sigma_0^{-1}\mu_0)$$

$\therefore E(\hat{\mu}_{MAP}) \neq \mu$

estimate is biased

(2) $$\hat{\Sigma}_{MAP} = \frac{1}{m}\sum_{i=1}^{m}[x^i x^{iT} - x^i\bar{x}^T - \bar{x}x^{iT} + \bar{x}\bar{x}^T]$$

$$= \frac{1}{m}\sum_{i=1}^{m}[x^i x^{iT}] - \left(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}\right)\bar{x}^T - \bar{x}\frac{1}{m}\sum_{i=1}^{m}x^{iT} + \bar{x}\bar{x}^T$$

$$= \frac{1}{m}\sum_{i=1}^{m}x^i x^{iT} - 2\bar{x}\bar{x}^T + \bar{x}\bar{x}^T$$

$$\hat{\Sigma}_{MAP} = \frac{1}{m}\sum_{i=1}^{m}x^i x^{iT} - \bar{x}\bar{x}^T \quad \text{—— (1)}$$

$x^i \sim N(\mu, \Sigma)$

$\bar{x} \sim N(\mu, \frac{1}{m}\Sigma)$ as $\bar{x} = \frac{1}{m}\sum_i x^i$

$$E(\bar{x}) = \frac{1}{m}\sum_{i=1} E(x^i) = \frac{m}{m}\mu = \mu$$

$$\begin{aligned}
Cov(\bar{x}) &= E(\bar{x}-\mu)(\bar{x}-\mu)^T \\
&= E\left(\frac{\sum x^i}{m}-\mu\right)\left(\frac{\sum x^i}{m}-\mu\right)^T \\
&= \frac{1}{m^2}E\left(\sum_{i=1}^{m} x^i - m\mu\right)\left(\sum_{i=1}^{m} x^i - m\mu\right)^T \\
&= \frac{1}{m^2}E\left(\sum_{i=1}^{m}(x^i-\mu)\right)\left(\sum_{i=1}^{m}(x^i-\mu)\right)^T \\
&= \frac{1}{m^2}\left[\sum_{i=1}^{m} E(x^i-\mu)(x^i-\mu)^T + \underbrace{\sum_i\sum_j E(x^i-\mu)(x^j-\mu)^T}_{=0 \text{ since } x^i\text{'s are independent}}\right] \\
&= \frac{1}{m^{\cancel{2}}}\cancel{m}\,\Sigma \\
&= \frac{1}{m}\Sigma
\end{aligned}$$

now, from ①

$$\begin{aligned}
E(\hat{\Sigma}_{MAP}) &= E\left(\frac{1}{m}\sum_{i=1}^{m} x^i x^{iT} - \bar{x}\bar{x}^T\right) \\
&= \frac{1}{m}\sum_{i=1}^{m} E(x^i x^{iT}) - E(\bar{x}\bar{x}^T)
\end{aligned}$$

(using linearity of $E()$ f^n)

$$\begin{aligned}
&= \frac{1}{m}\sum_{i=1}^{m}(\Sigma + \mu\mu^T) - \left(\frac{1}{m}\Sigma + \mu\mu^T\right) \\
&= \Sigma + \cancel{\mu\mu^T} - \frac{1}{m}\Sigma - \cancel{\mu\mu^T} \\
&= \frac{m-1}{m}\Sigma
\end{aligned}$$

$\therefore$ $\hat{\Sigma}_{MAP}$ is biased estimator

So is $\hat{\Sigma}_{MLE}$