2. Mathematical Foundamentals.

2.1 Probability.

1. $$E\|x\|_2^2 = \sum_{i=1}^{d} E(X_i^2)$$

$$= \sum_{i=1}^{d} E^2(X_i) + Var(X_i)$$

$$= \sum_{i=1}^{d} \mu_i^2 + \sigma_{i,i}^2$$

$$E\|x-y\|_2^2 = \sum_{i=1}^{d} E(X_i - Y_i)^2$$

$$= \sum_{i=1}^{d} 2[E(X_i^2) - E(X_i \cdot Y_i)]$$

$$= \sum_{i=1}^{d} 2[E(X_i)^2 + Var(X_i) - E(X_i) \cdot E(Y_i)]$$

$$= 2\sum_{i=1}^{d} Var(X_i)$$

$$= 2 \cdot \sum_{i=1}^{d} \sigma_{i,i}^2$$

2. $$P(Z \le z) = P(\alpha_i X_i + \alpha_j X_j \le z)$$

$$= \int_{X_k, k \ne i,j} \int_{\alpha_i X_i + \alpha_j X_j \le z} f(\vec{x}) d\vec{x}$$

$$= \int_{\alpha_i X_i + \alpha_j X_j \le z} f(x_i, x_j) dx_i dx_j.$$

$$f(X_i, X_j) \sim N(\mu', \Sigma'),\ \mu' = (\mu_i, \mu_j),\ \Sigma' = \begin{pmatrix} \pi_{i,i} & \pi_{i,j} \\ \pi_{i,j} & \pi_{j,j} \end{pmatrix}$$

Similar to 3., $f_Z(z) \sim N(\mu', \sigma')$

$$\mu' = \alpha_i \mu_i + \alpha_j \mu_j,\ \sigma' = \sqrt{\alpha_i^2 \pi_{i,i} + \alpha_j^2 \pi_{j,j} + 2\alpha_i \alpha_j \pi_{i,j}}$$

3. $f_W(w) \sim N(\mu_1, \sigma_1)$, $f_R(r) \sim N(\mu_2, \sigma_2)$.

Let $Z = W + R$. 则 $P(Z \le z) = P(W + R \le z)$

$$= \iint_{w+r<z} f_W(w) \cdot f_R(r) dw dr$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \iint_{w+r\le z} \exp\left\{-\left[\frac{(w-\mu_1)^2}{2\sigma_1^2} + \frac{(r-\mu_2)^2}{2\sigma_2^2}\right]\right\} dw dr$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \iint_{w+r\le z-\mu_1-\mu_2} \exp\left[-\left(\frac{w^2}{2\sigma_1^2} + \frac{r^2}{2\sigma_2^2}\right)\right] dw dr$$

Let $z = w + r$, $$P(Z \le z) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{z-\mu_1-\mu_2} \int_{-\infty}^{+\infty} e^{-\left[\frac{(z-r)^2}{\sigma_1^2} + \frac{r^2}{\sigma_2^2}\right]/2} dr dz$$

$$= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{z-\mu_1-\mu_2}\int_{-\infty}^{+\infty} \exp\left[-\left(\frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}r^2 - \frac{2}{\sigma_1^2}zr + \frac{z^2}{\sigma_1^2}\right)/2\right]$$

$$= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{z-\mu_1-\mu_2}\int_{-\infty}^{+\infty} \exp\left\{\left[-\left(\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1\sigma_2}r - \frac{\sigma_2 z}{\sqrt{\sigma_1^2+\sigma_2^2}\,\sigma_1}\right)^2 + \frac{z^2}{\sigma_1^2+\sigma_2^2}\right]/2\right\}drdz$$

$$= \frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{z-\mu_1-\mu_2} \exp\left[\frac{-z^2}{2(\sigma_1^2+\sigma_2^2)}\right]dz \cdot \int_{-\infty}^{+\infty} \exp\left[-r^2/\left(2\cdot\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)\right]$$

$$= \frac{1}{2\pi\sigma_1\sigma_2}\cdot\sqrt{2\pi}\cdot\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}\int_{-\infty}^{z-\mu_1-\mu_2} \exp\left[\frac{-z^2}{2(\sigma_1^2+\sigma_2^2)}\right]dz$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\int_{-\infty}^{z-\mu_1-\mu_2} \exp\left[\frac{-z^2}{2(\sigma_1^2+\sigma_2^2)}\right]dz$$

$$f(z) = \frac{d}{dz}P(Z\le z) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\exp\left[-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right].$$

Therefore, $w+R$ is subjected to Gaussian distribution.

2.2 Linear Algebra

1. The dimension of S_A is $d-k$.

2. $\vec{w}$ can be decomposed into linear combination of $\vec{v}_1, \vec{v}_2, \cdots, \vec{v}_k$ and one vector orthogonal the subspace S_v. That is, $\vec{w} = \sum_{i=1}^{k}\alpha_i\cdot\vec{v}_i + \vec{u}$

$\vec{x}\in S_v$ so $\vec{x} = \sum_{i=1}^{k}\beta_i\vec{v}_i$, $\vec{w}-\vec{x} = \sum_{i=1}^{k}(\alpha_i-\beta_i)\vec{v}_i + \vec{u}$

$$\|w-x\|_2 = \sqrt{(\vec{w}-\vec{x})^T\cdot(\vec{w}-\vec{x})} = \sum_{i=1}^{k}(\alpha_i-\beta_i)^2\vec{v}_i^T\cdot\vec{v}_i + \vec{u}^T\cdot\vec{u} \geq \vec{u}^T\cdot\vec{u}.$$

Hence $\vec{x}^* = \sum_{i=1}^{k}\alpha_i\vec{v}_i$, $\alpha_i = \vec{w}^T\cdot\vec{v}_i/\vec{v}_i^T\cdot\vec{v}_i$

3. $$x^* = \sum_{i=1}^{k}\frac{w^Tv_i\cdot v_i}{v_i^T\cdot v_i}$$

$$= (v_1, v_2, \cdots, v_k)\cdot\begin{pmatrix} v_1^T\cdot w/v_1^T\cdot v_1 \\ \vdots \\ v_k^T\cdot w/v_k^T\cdot v_k\end{pmatrix}$$

$$= (v_1, \cdots, v_k) \begin{pmatrix} v_1^T / v_1^T \cdot v_1 \\ \vdots \\ v_k^T / v_k^T \cdot v_k \end{pmatrix} \cdot w$$

$$= \sum_{i=1}^{k} v_i \cdot v_i^T / (v_i^T \cdot v_i) \cdot w$$

So $M = \sum_{i=1}^{k} \frac{v_i \cdot v_i^T}{v_i^T \cdot v_i}$

3. Linear Regression

3.1 Feature Normalization.

The modified code is in the workspace directory.

3.2 Gradient Descent Setup.

1. $J(\theta) = \frac{1}{m} (X \cdot \theta - y)^T (X\theta - y)$

2. $\nabla_\theta J(\theta) = 2 (X^T X \theta - X^T y) / m$

3. The first-order approximization of $J(\theta + \eta h) - J(\theta)$ should be $\nabla_\theta^T J(\theta) \cdot \eta h$

4. For GD algorithm, $h = -\nabla_\theta J(\theta)$. $\theta \leftarrow \theta - 2\eta (X^T X \theta - X^T y)/m$

5.6 The modified code is in the workspace directory.

3.3 (Optional) Gradient Checker

3.4 Batch Gradient Descent.

Ps: Both these two above sections are code completion

3.5 Ridge Regression

1. $\nabla_\theta J(\theta) = \frac{2}{m} J_\theta(h) \cdot [h_\theta(X) - y] + 2\lambda\theta$, $h_\theta(X) = \begin{pmatrix} h_\theta(x_1) \\ \vdots \\ h_\theta(x_m) \end{pmatrix}$

$J(\theta) = \frac{1}{m}(h_\theta^T(X) - y^T)(h_\theta(X) - y) + \lambda\theta^T\theta$

Known that $dh_\theta(X) = J_\theta(h) \cdot d\theta$,

$dJ(\theta) = \frac{1}{m}[d\theta^T J_\theta^T(h) + h_\theta^T(X) - y^T][J_\theta(h)d\theta + h_\theta(X) - y] + \lambda(\theta^T + d\theta^T)(\theta + d\theta)$

$= \frac{2}{m} d\theta^T \cdot J_\theta^T(h)[h_\theta(X) - y] + 2\lambda d\theta^T \cdot \theta$

Hence $\nabla_\theta J(\theta) = \frac{2}{m} J_\theta^T(h)[h_\theta(X) - y] + 2\lambda\theta$

2.3: Code completion.

4. If bias term B is large enough, the parameter with respect to B, θ_B, can be small enough to be omitted in regularization. When $\theta_B^2 / \|\theta\|_2^2$ is very as small as you like, the regularization on B is weak enough.

5. (Optional)

Statement: When B goes to infinity, the regularization on the bias term disappear.

Proof:

$h_\theta(x_i) = h_{\theta'}(x_i') + \theta_B \cdot B$. θ' is θ excluding θ_B, x_i' is x_i excluding bias term.

So $J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_{\theta'}(x_i') + \theta_B \cdot B - y_i)^2 + \lambda(\theta'^T\theta' + \theta_B^2)$

When $B \to \infty$, if $\theta_B \neq 0$, $J(\theta) = +\infty$. Hence $\lim_{B\to\infty} \theta_B = 0$.

6.(optional). Not finished yet.

7.8 Not finished yet

3.6 Stochastic Gradient Descent

1. $f_i(\theta) = (h_\theta(x_i) - y_i)^2 + \lambda \cdot \theta^T \theta$

2. This conclusion has been proved in SGD's concept check questions.

3. $\theta \leftarrow \theta - 2\eta[(h_\theta(x_i) - y_i) \cdot \nabla_\theta h_\theta(x_i) + \lambda\theta]$, η is step size.

4.5.6: Not finished yet.

4. Risk Minimization.

4.1 Square Loss.

1. This has been proved in ESR's concept check question 2(b).

2.(a) $f^*(x) = \underset{a}{\arg\min}\, E[(a-Y)^2 | X=x]$

$E[(a-Y)^2|X] = a^2 - 2aE(Y|X) + E(Y^2|X)$

$= a^2 - 2aE(Y|X) + E^2(Y|X) + Var(Y|X)$

$= (a - E(Y|X))^2 + Var(Y|X)$

So $f^*(x) = E(Y|X=x)$

(b) $E[(f(X)-Y)^2] = E\{E[(f(X)-Y)^2|X]\}$

Known that: $E[(f^*(X)-Y)^2|X] \le E[(f(X)-Y)^2|X]$

So $E\{E[(f^*(X)-Y)^2|X]\} \le E\{E[(f(X)-Y)^2|X]\}$

4.2 (Optional) Median Loss.

1. This has been proved in Intro's concept check questions.