L1 Supervised Learning (Jesus Please finish the basics)

1. Linear Regression
   a. Model

$$h(x) = \sum_{i=0}^{n} w_i x_i = w^T x$$

   b. Cost func (OLS)

$$\mathcal{L}(w) = \frac{1}{2}\sum_{i=1}^{m}\left(h\left(x^{(i)}\right) - y^{(i)}\right)^2$$

   c. Gradient Descent

$$w_j \coloneqq w_j - \alpha \frac{\partial \mathcal{L}(w)}{\partial w_j}$$

   d. $\frac{\partial \mathcal{L}(w)}{\partial w_j}$

$$\frac{\partial \mathcal{L}(w)}{\partial w_j} = (h(x) - y)x_j$$

   - Derivation

$$\begin{aligned}\frac{\partial \mathcal{L}(w)}{\partial w_j} &= \frac{\partial}{\partial w_j}\left(\frac{1}{2}(h(x) - y)^2\right) \\ &= (h(x) - y) \cdot \frac{\partial}{\partial w_j}(h(x) - y) \\ &= (h(x) - y) \cdot x_j\end{aligned}$$

   e. Batch GD (LMS) (each step)

$$w_j \coloneqq w_j - \alpha \sum_{i=1}^{m}\left(h(x)^{(i)} - y^{(i)}\right)x_j^{(i)}$$

   f. Stochastic GD (each step)

$$w_j \coloneqq w_j - \alpha\left(h(x)^{(i)} - y^{(i)}\right)x_j^{(i)}$$

   g. Matrix Derivatives
   - Gradient

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}$$

   - Trace

1) Formula

$$\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ii}$$

2) Properties

$$\begin{aligned} \mathrm{tr}(ABC) &= \mathrm{tr}(CAB) = \mathrm{tr}(BCA) \\ \mathrm{tr}(A) &= \mathrm{tr}(A^T) \\ \mathrm{tr}(A+B) &= \mathrm{tr}(A) + \mathrm{tr}(B) \\ \mathrm{tr}(\alpha A) &= \alpha \mathrm{tr}(A) \end{aligned}$$

3) Combined Properties

$$\begin{aligned} \nabla_A \mathrm{tr}(AB) &= B^T \\ \nabla_{A^T} f(A) &= \left(\nabla_A f(A)\right)^T \\ \nabla_A \mathrm{tr}(ABA^TC) &= CAB + C^TAB^T \\ \nabla_A |A| &= |A|(A^{-1})^T \end{aligned}$$

h. Normal Equation

- Formula

$$\theta = (X^TX)^{-1}X^Ty$$

- Derivation

$$\begin{aligned} \nabla_w \mathcal{L}(w) &= \nabla_w \frac{1}{2}(Xw-y)^T(Xw-y) \\ &= \frac{1}{2}\nabla_w(w^TX^TXw - w^TX^Ty - y^TXw + y^Ty) \\ &= \frac{1}{2}\nabla_w \mathrm{tr}(w^TX^TXw - w^TX^Ty - y^TXw + y^Ty) \\ &= \frac{1}{2}\nabla_w\left(\mathrm{tr}(w^TX^TXw) - 2\mathrm{tr}(y^TXw)\right) \\ &= \frac{1}{2}(2X^TXw - 2X^Ty) \\ &= X^TXw - X^Ty \\ &\Longrightarrow \theta = (X^TX)^{-1}X^Ty \end{aligned}$$

i. Probabilistic Interpretation

- Model

$$y^{(i)} = w^Tx^{(i)} + \epsilon^{(i)}$$

1) $\epsilon^{(i)}: i.i.d$ + Gaussian

$$p\left(\epsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(\epsilon^{(i)}\right)^2}{2\sigma^2}}$$

- Probabilistic Model

$$p\left(y^{(i)}\middle|x^{(i)},w\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\left(y^{(i)}-w^Tx^{(i)}\right)^2}{2\sigma^2}}$$

- Likelihood Func

$$\begin{aligned}L(w)&=\prod_{i=1}^{m}p\left(y^{(i)}\middle|x^{(i)},w\right)\\&=\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\left(y^{(i)}-w^Tx^{(i)}\right)^2}{2\sigma^2}}\end{aligned}$$

- MLE

$$\begin{aligned}\ell(w)&=\log\ell(w)\\&=\log\prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\left(y^{(i)}-w^Tx^{(i)}\right)^2}{2\sigma^2}}\\&=\sum_{i=1}^{m}\log\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\left(y^{(i)}-w^Tx^{(i)}\right)^2}{2\sigma^2}}\\&=m\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(y^{(i)}-w^Tx^{(i)}\right)^2\end{aligned}$$

$$\begin{aligned}\underset{w}{\operatorname{argmax}}\,\ell(w)&=\underset{w}{\operatorname{argmax}}\,m\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(y^{(i)}-w^Tx^{(i)}\right)^2\\&=\underset{w}{\operatorname{argmax}}\left(-\sum_{i=1}^{m}\left(y^{(i)}-w^Tx^{(i)}\right)^2\right)\\&=\underset{w}{\operatorname{argmin}}\sum_{i=1}^{m}\left(y^{(i)}-w^Tx^{(i)}\right)^2\end{aligned}$$

j. Locally Weighted Linear Regression
- Original Linear Regression
  1) S1: $w\leftarrow\underset{w}{\operatorname{argmin}}\sum_{i=1}^{m}\left(y^{(i)}-w^Tx^{(i)}\right)^2$
  2) S2: $\hat{y}\leftarrow w^Tx$
- Weighted Linear Regression
  1) S1: $w\leftarrow\underset{w}{\operatorname{argmin}}\sum_{i=1}^{m}e^{-\frac{\left(x^{(i)}-x\right)^2}{2\tau^2}}\cdot\left(y^{(i)}-w^Tx^{(i)}\right)^2$
     a. τ: bandwidth param
     b. $\left|x^{(i)}-x\right|$
        - Small → weight ≈ 1
        - Large → weight ≈ 0
  2) S2: $\hat{y}\leftarrow w^Tx$

2. Classification & Logistic Regression
   a. LogReg

- Formula

$$h(x) = g(w^T x)$$

- Sigmoid func

$$g(z) = \frac{1}{1 + e^{-z}}$$

1) Derivative

$$g'(z) = g(z)\big(1 - g(z)\big)$$

a. Derivation

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1 + e^{-z}} \\ &= \frac{e^{-z} + 1 - 1}{(1 + e^{-z})^2} \\ &= \frac{1}{1 + e^{-z}}\left(1 - \frac{1}{1 + e^{-z}}\right) \\ &= g(z)\big(1 - g(z)\big) \end{aligned}$$

- MLE

1) Assumptions

$$P(y = 1|x, w) = h(x)$$
$$P(y = 0|x, w) = 1 - h(x)$$

2) Prob Model

$$p(y|x, w) = h(x)^y \big(1 - h(x)\big)^{1-y}$$

3) Likelihood Func

$$L(w) = \prod_{i=1}^{m} h\big(x^{(i)}\big)^{y^{(i)}} \left(1 - h\big(x^{(i)}\big)\right)^{1-y^{(i)}}$$

4) Log likelihood

$$\ell(w) = \sum_{i=1}^{m} y^{(i)} \log h\big(x^{(i)}\big) + \big(1 - y^{(i)}\big) \log\left(1 - h\big(x^{(i)}\big)\right)$$

5) MLE(SGD)

$$\begin{aligned} \frac{\partial \ell(w)}{\partial w_j} &= \left(\frac{y}{g(w^T x)} - \frac{1 - y}{1 - g(w^T x)}\right)\frac{\partial g(w^T x)}{\partial w_j} \\ &= \left(\frac{y}{g(w^T x)} - \frac{1 - y}{1 - g(w^T x)}\right) g(w^T x)\big(1 - g(w^T x)\big)\frac{\partial (w^T x)}{\partial w_j} \\ &= \Big(y\big(1 - g(w^T x)\big) - (1 - y) g(w^T x)\Big) x_j \\ &= \big(y - h(x)\big) x_j \end{aligned}$$

6) SGA

$$w_j := w_j + \alpha \left(y^{(i)} - h\left(x^{(i)}\right) \right) x_j^{(i)}$$

b. GD - Newton's Method

$$w := w - \frac{f(w)}{f'(w)}$$

- Newton-Raphson Method

$$w := w - H^{-1} \nabla_w \ell(w)$$

- Hessian

$$H_{ij} = \frac{\partial^2 \ell(w)}{\partial w_i \partial w_j}$$

- Newton VS GD
  1) Adv: faster convergence, fewer iterations
  2) Disadv: one iteration is expensive as hell (we need to find & invert a $n \times n$ Hessian)

3. Generalized Linear Models (GLM)
   a. Summary
      - Regression: $y|x,\theta \sim N(\mu, \sigma^2)$
      - Classification: $y|x,\theta \sim Bernoulli(\phi)$
      - Regression + Classification → Generalized Linear Models (GLMs)

   b. Exponential family
      - General form of distribution

$$p(y, \eta) = b(y) \cdot e^{\eta^T T(y) - a(\eta)}$$

      - Notations
        1) η: natural parameter (i.e. canonical parameter)
        2) $T(y)$: sufficient statistic (usually $T(y) = y$)
        3) $a(\eta)$: log partition function
        4) $e^{-a(\eta)}$: normalization constant (ensure that $\int p(y, \eta) dy = 1$)
        5) $T, a, b =$ fixed: fixed choice that defines a family/set of distributions that is parametrized by η
           a. As we vary η, we then get different distributions within this family
      - e.g. Bernoulli Distribution

$$\begin{aligned} p(y, \phi) &= \phi^y (1 - \phi)^{1-y} \\ &= e^{y \log \phi + (1-y) \log(1-\phi)} \\ &= e^{\log \frac{\phi}{1-\phi} \cdot y + \log(1-\phi)} \end{aligned}$$

        1) $T(y) = y$
        2) $a(\eta) = \log(1 + e^{\eta})$
        3) $b(y) = 1$

4) $\eta = \log\frac{\phi}{1-\phi} \Leftrightarrow \phi = \frac{1}{1+e^{-\eta}}$

- e.g. Gaussian Distribution (set $\sigma^2 = 1$)

$$p(y,\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}$$
$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \cdot e^{\mu y - \frac{1}{2}\mu^2}$$

1) $T(y) = y$
2) $a(\eta) = \frac{\eta^2}{2}$
3) $b(\eta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$
4) $\eta = \mu$

- Other members of Exponential Family:
1) Multinomial
2) Poisson: modeling count-data
3) Gamma: modeling continuous, non-negative random vars (e.g. time intervals)
4) Beta & Dirichlet: distributions over probabilities

4. Constructing GLMs
   a. 3 assumptions
      - $y|x,\theta \sim \text{ExponentialFamily}(\eta)$
      - $h(x) = E[y|x]$
      - $\eta_i = w_i^T x$ ($\eta \sim x$ linearly)
   b. Derivation examples
      - e.g. OLS

$$h(x) = E[y|x,w] = \mu = \eta = w^T x$$

- e.g. LogReg

$$h(x) = E[y|x,w] = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-w^T x}}$$

- e.g. Softmax Regression
1) Problem setting

$$y \in \{1, \dots, k\}$$

2) Define parameters: $\phi_1, \dots, \phi_{k-1}$

$$\phi_i = p(y = i, \phi)$$

$$\phi_k = p(y = k, \phi) = 1 - \sum_{i=1}^{k-1} \phi_i$$

3) Linear Transformation (encoding): $T(y) \in \mathbb{R}^{k-1}$

$$T(1)=\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}, T(2)=\begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix}, \dots, T(k-1)=\begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix}, T(k)=\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}$$

4) Indicator Function: $I\{\text{True}\}=1, I\{\text{False}\}=0$

$$T(y)_i = I\{y=i\}$$

$$E[T(y)_i] = P(y=i) = \phi_i$$

$$\sum I\{y=i\} = 1$$

5) Derivation of Exponential Family

$$\begin{aligned}
p(y,\phi) &= \prod_{i=1}^{k} \phi_i^{I\{y=i\}} \\
&= \prod_{i=1}^{k-1} \phi_i^{I\{y=i\}} \cdot \phi_k^{1-\sum_{i=1}^{k-1} I\{y=i\}} \\
&= \prod_{i=1}^{k-1} \phi_i^{T(y)_i} \cdot \phi_k^{1-\sum_{i=1}^{k-1} T(y)_i} \\
&= e^{\log \prod_{i=1}^{k-1} \phi_i^{T(y)_i} \cdot \phi_k^{1-\sum_{i=1}^{k-1} T(y)_i}} \\
&= e^{\left(\sum_{i=1}^{k-1} T(y)_i \log \phi_i\right) + \left(1-\sum_{i=1}^{k-1} T(y)_i\right) \log \phi_k} \\
&= e^{\left(\sum_{i=1}^{k-1} T(y)_i \log \phi_i\right) - \left(\sum_{i=1}^{k-1} T(y)_i \log \phi_k\right) + \log \phi_k} \\
&= e^{\left(\sum_{i=1}^{k-1} T(y)_i \log \frac{\phi_i}{\phi_k}\right) + \log \phi_k} \\
&= b(y) e^{\eta^T T(y) - a(\eta)}
\end{aligned}$$

a. $T(y) = T(y) = \begin{bmatrix} I\{y=1\} \\ I\{y=2\} \\ \vdots \\ I\{y=k-1\} \end{bmatrix}$

b. $\eta = \begin{bmatrix} \log \frac{\phi_1}{\phi_k} \\ \log \frac{\phi_2}{\phi_k} \\ \vdots \\ \log \frac{\phi_{k-1}}{\phi_k} \end{bmatrix}$

c. $a(\eta) = -\log \phi_k$

d. $b(y) = 1$

6) Derivation of Distribution

a. Derive ϕ_i

$$\eta_i = \log \frac{\phi_i}{\phi_k}$$

$$
\begin{aligned}
e^{\eta_i} &= \frac{\phi_i}{\phi_k} \\
\phi_k e^{\eta_i} &= \phi_i \\
\phi_k \sum_{i=1}^{k} e^{\eta_i} &= \sum_{i=1}^{k} \phi_i = 1 \\
\phi_k &= \frac{1}{\sum_{i=1}^{k} e^{\eta_i}} \\
\phi_i &= \phi_k e^{\eta_i} = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}
\end{aligned}
$$

b. Derive distribution

$$
\begin{aligned}
p(y = i|x, w) &= \phi_i \\
&= \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}} \\
&= \frac{e^{w_i^T x}}{\sum_{j=1}^{k} e^{w_j^T x}}
\end{aligned}
$$

7) Hypothesis

$$
\begin{aligned}
h(x) &= E[T(y)|x, w] \\
&= E\begin{bmatrix} \begin{matrix} I\{y = 1\} \\ I\{y = 2\} \\ \vdots \\ I\{y = k-1\} \end{matrix} & |x, w \end{bmatrix} \\
&= \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_{k-1} \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{e^{w_1^T x}}{\sum_{j=1}^{k} e^{w_j^T x}} \\ \vdots \\ \dfrac{e^{w_{k-1}^T x}}{\sum_{j=1}^{k} e^{w_j^T x}} \end{bmatrix}
\end{aligned}
$$

8) Log likelihood

$$
\begin{aligned}
\ell(w) &= \sum_{i=1}^{m} \log p\left(y^{(i)} \middle| x^{(i)}, w\right) \\
&= \sum_{i=1}^{m} \log \prod_{h=1}^{k} \left(\frac{e^{w_i^T x^{(i)}}}{\sum_{j=1}^{k} e^{w_j^T x^{(i)}}} \right)^{I\{y^{(i)} = h\}}
\end{aligned}
$$

L2 Generative Learning Algorithms

1. Bayes Summary
    a.
2. Gaussian Discriminant Analysis
3. Naïve Bayes