CS229 Machine Learning Notes and Formulae

Chang He

July 2019

Contents

1 Supervised Learning and Linear Models

1.1 Linear Models

1.1.1 Linear Regression

Given data set X with n data points and d features, with the associated targets y, minimize the loss function:

$$\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\theta}^T X_{(i)} + b - y_{(i)})^2$$

By changing the definition of X and $\boldsymbol{\theta}$ to include the intercept term b, with the 0th feature of X being 1 and 0th term of $\boldsymbol{\theta}$ being b, we have:

$$\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\theta}^T X_{(i)} - y_{(i)})^2$$

Vectorizing the loss function, we have the objective function:

$$\min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \frac{1}{2}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y})$$

1.1.1.1 Ordinary Least Square

To find the potential minimum of this function, we find its gradient first:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}}\, \frac{1}{2}(X\boldsymbol{\theta} - \boldsymbol{y})^T(X\boldsymbol{\theta} - \boldsymbol{y}) \\
&= \nabla_{\boldsymbol{\theta}}\, \frac{1}{2}(\boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - \boldsymbol{\theta}^T X^T \boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\theta} + \boldsymbol{y}^T\boldsymbol{y}) \\
&= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\, \text{tr}\, (\boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - \boldsymbol{\theta}^T X^T \boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\theta} + \boldsymbol{y}^T\boldsymbol{y}) && \text{tr } x = x \text{ if } x \in \mathcal{R} \\
&= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\, (\text{tr}\, \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - \text{tr}\, \boldsymbol{\theta}^T X^T \boldsymbol{y} - \text{tr}\, \boldsymbol{y}^T X\boldsymbol{\theta}) && \text{remove the constant} \\
&= \frac{1}{2}\nabla_{\boldsymbol{\theta}}\, (\text{tr}\, \boldsymbol{\theta}^T X^T X\boldsymbol{\theta} - \text{tr}\, \boldsymbol{y}^T X\boldsymbol{\theta} - \text{tr}\, \boldsymbol{y}^T X\boldsymbol{\theta}) && \text{tr } A^T = \text{tr } A \\
&= \frac{1}{2}(\nabla_{\boldsymbol{\theta}^T}\, \text{tr}\, \boldsymbol{\theta}^T X^T X\boldsymbol{\theta})^T - \nabla_{\boldsymbol{\theta}}\, \text{tr}\, \boldsymbol{y}^T X\boldsymbol{\theta} && \nabla_A\, f(A) = (\nabla_{A^T}\, f(A))^T \\
&= \frac{1}{2}(\boldsymbol{\theta}^T X^T X + \boldsymbol{\theta}^T X^T X) - \boldsymbol{y}^T X && \nabla_A\, ABA^T C = CAB + C^T AB^T \\
& && \text{with } A = \boldsymbol{\theta}^T,\ B = X^T X,\ C = I \\
&= (\boldsymbol{\theta}^T X^T - y^T)X \\
&= (X\boldsymbol{\theta} - y)^T X \\
&= X^T(X\boldsymbol{\theta} - y)
\end{aligned}$$

To obtain the critical point, set the gradient to 0 and solve the linear equation:

$$X^T X\boldsymbol{\theta} = X^T\boldsymbol{y}$$

The result is denoted as $\theta = (X^T X)^{-1}X^T\boldsymbol{y}$.

1.1.1.2 Probabilistic interpretation of Linear Model

Given data set X with n data points and d features, with the associated targets $\boldsymbol{y}$, assuming that the data set is modeled by parameter $\boldsymbol{\theta}$ and the residual $(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})$ follows a Gaussian distribution, we have:

$$\begin{aligned}
\boldsymbol{y}_{(i)} &= \boldsymbol{\theta}^T X_{(i)} + \boldsymbol{\epsilon}_{(i)} && \boldsymbol{\epsilon}_{(i)} \sim \mathcal{N}(0, \sigma^2) \\
P(\boldsymbol{\epsilon}_{(i)}) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\boldsymbol{\epsilon}_{(i)})^2}{2\sigma^2}\right) \\
P(\boldsymbol{y}_{(i)}|X_{(i)};\boldsymbol{\theta}) &= \mathcal{N}(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)}, \sigma^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})^2}{2\sigma^2}\right)
\end{aligned}$$

Assuming $\boldsymbol{\epsilon}_{(i)}$ is Independently and Identically Distributed (IDD), we have:

$$\begin{aligned} P(\boldsymbol{y}|X;\boldsymbol{\theta}) &= \prod_{i=1}^{n} P(\boldsymbol{y}_{(i)}|X_{(i)};\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

Let find the maximum likelihood estimate of $\boldsymbol{\theta}$ by maximizing $P(\boldsymbol{y}|X;\boldsymbol{\theta})$:

$$\begin{aligned} \max_{\boldsymbol{\theta}} P(\boldsymbol{y}|X;\boldsymbol{\theta}) &= \max_{\boldsymbol{\theta}} \ln P(\boldsymbol{y}|X;\boldsymbol{\theta}) && \text{since ln is monotonously increasing} \\ &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})^2}{2\sigma^2}\right)\right) \\ &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})^2}{2\sigma^2}\right) \\ &= \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(\frac{(\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})^2}{2\sigma^2}\right) && \text{remove the constant term} \\ &= \min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{y}_{(i)} - \boldsymbol{\theta}^T X_{(i)})^2 \\ &= \min_{\boldsymbol{\theta}} \frac{1}{2} (X\boldsymbol{\theta} - \boldsymbol{y})^T (X\boldsymbol{\theta} - \boldsymbol{y}) \\ &= \min_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \end{aligned}$$

1.1.2 Linear Models for Classification

1.1.2.1 Using Linear Regression for Classification Since the component range of hypothesis function $h_{\boldsymbol{\theta}}(X) = X\boldsymbol{\theta}$ lies far outside of the relevant range for classification $[0, 1]$, linear regression is extremely sensitive to data points with a large norm, so it is generally not recommended to use linear regression for classification.

1.1.2.2 Logistic Regression The hypothesis for logistic regression is:

$$h_{\boldsymbol{\theta}}(X) = S(\boldsymbol{\theta}^T X) = \frac{1}{1 + \exp(-X\boldsymbol{\theta})}$$

, where function S is called sigmoid or logistic function.

Each component of the hypothesis, $h_{\boldsymbol{\theta}}(X_{(i)})$, represents the probability of the associated target $\boldsymbol{y}_{(i)}$ equal to 1, that is:

$$\begin{aligned} h_{\boldsymbol{\theta}}(X_{(i)}) &= P(\boldsymbol{y}_{(i)} = 1|X_{(i)};\boldsymbol{\theta}) \\ 1 - h_{\boldsymbol{\theta}}(X_{(i)}) &= P(\boldsymbol{y}_{(i)} = 0|X_{(i)};\boldsymbol{\theta}) \end{aligned}$$

Since $y \in \{0, 1\}$,

$$P(\boldsymbol{y}_{(i)}|X_{(i)};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(X_{(i)})^y (1 - h_{\boldsymbol{\theta}}(X_{(i)}))^{(1-y)}$$

Using MLE estimation, with IDD assumption, we get:

$$\begin{aligned} &\max_{\boldsymbol{\theta}} \ln P(\boldsymbol{y}|X;\boldsymbol{\theta}) \\ &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln P(\boldsymbol{y}_{(i)}|X_{(i)};\boldsymbol{\theta}) \\ &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln\left(h_{\boldsymbol{\theta}}(X_{(i)})^y (1 - h_{\boldsymbol{\theta}}(X_{(i)}))^{(1-y)}\right) \\ &= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y \ln(h_{\boldsymbol{\theta}}(X_{(i)})) + (1 - \boldsymbol{y}_{(i)}) \ln(1 - h_{\boldsymbol{\theta}}(X_{(i)}))) \end{aligned}$$

Given that $1 - S(x) = \frac{\exp(-x)}{1+\exp(-x)} = \exp(-x)S(x)$

$$
\begin{aligned}
&= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(\boldsymbol{y}_{(i)} \ln(h_{\boldsymbol{\theta}}(X_{(i)})) + (1 - \boldsymbol{y}_{(i)})(-\boldsymbol{\theta}^T X_{(i)} + \ln(h_{\boldsymbol{\theta}}(X_{(i)}))) \Big) \\
&= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(\boldsymbol{y}_{(i)} \ln(h_{\boldsymbol{\theta}}(X_{(i)})) + (-\boldsymbol{\theta}^T X_{(i)} + \ln(h_{\boldsymbol{\theta}}(X_{(i)}))) + (\boldsymbol{\theta}^T X_{(i)} \boldsymbol{y}_{(i)} - \boldsymbol{y}_{(i)} \ln(h_{\boldsymbol{\theta}}(X_{(i)}))) \Big) \\
&= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(\ln(h_{\boldsymbol{\theta}}(X_{(i)})) - (1 - \boldsymbol{y}_{(i)}) \boldsymbol{\theta}^T X_{(i)} \Big) \\
&= \min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(\ln(1 + \exp(-\boldsymbol{\theta}^T X_{(i)})) + (1 - \boldsymbol{y}_{(i)}) \boldsymbol{\theta}^T X_{(i)} \Big)
\end{aligned}
$$

The gradient of which is:

$$
\begin{aligned}
&\nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(\ln(1 + \exp(-\boldsymbol{\theta}^T X_{(i)})) + (1 - \boldsymbol{y}_{(i)}) \boldsymbol{\theta}^T X_{(i)} \Big) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{1 + \exp(\boldsymbol{\theta}^T X_{(i)})} \circ X_{(i)} + (1 - \boldsymbol{y}_{(i)}) X_{(i)} \right) \\
&= \sum_{i=1}^{n} \left(1 - \boldsymbol{y}_{(i)} - \frac{1}{1 + \exp(\boldsymbol{\theta}^T X_{(i)})} \right) \circ X_{(i)} \\
&= \sum_{i=1}^{n} \left(\frac{1}{1 + \exp(-\boldsymbol{\theta}^T X_{(i)})} - \boldsymbol{y}_{(i)} \right) \circ X_{(i)} \\
&= (h(X_{(i)}) - \boldsymbol{y}_{(i)}) \circ X_{(i)}
\end{aligned}
$$

1.2 Newton's Method

$$\boldsymbol{\theta} := \boldsymbol{\theta} - (\nabla^2_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta}))^{-1} \nabla_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta})$$

Newton's Method enjoys quadratic convergence, meaning that the error of the result will decrease quadratically. However, it is more expensive, both time and memory-wise, to calculate and invert Hessian in each iteration compared to gradient descent. For data with a large number of dimensions, using Newton's Method is problematic also because of proliferation of saddle points, which Newton's Method tends to attract to.

Despite so, for linear models with mean-square loss, Newton's Method converges extremely fast. For a perfect square form loss function, Newton's Method converges in an exactly single iteration.

1.3 Generalized Linear Model

1.3.1 Exponential Family Distribution

Exponential Family Distribution is a class of distributions parameterized by η:

$$p(y; \eta) = b(y) \exp(\eta^T T(y) - a(\eta))$$

Both Bernoulli and Gaussian distribution are special cases of the exponential family distribution. For Bernoulli distribution, let's define:

$$
\begin{aligned}
p(y; \phi) &= \phi^y (1 - \phi)^{(1-y)} \\
&= \exp\big(y \ln \phi + (1 - y) \ln(1 - \phi)\big) \\
&= \exp\left(y \ln \frac{\phi}{1 - \phi} + \ln(1 - \phi)\right)
\end{aligned}
$$

In this case, we have:

$$b(y) = 1$$

$$
\begin{aligned}
\eta &= \frac{\phi}{1-\phi} \\
T(y) &= y \\
a(\eta) &= -\ln(1+\exp(\eta)) \\
p(y;\eta) &= \exp(\eta y - \ln(1+\exp(\eta))) \\
&= \frac{e^{\eta y}}{1+e^{\eta}} = \text{sigmoid}(\eta)\cdot\exp(\eta(y-1))
\end{aligned}
$$

For Gaussian distribution, we define:

$$
\begin{aligned}
p(y;\eta) &= b(y)\exp\left(\boldsymbol{\eta}^T T(y) - a(\boldsymbol{\eta})\right) \\
p(y;(\mu,\sigma)) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{2y\mu - y^2}{2\sigma^2}\right)\exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}}\sigma^{-1}\exp\left(y\cdot\frac{\mu}{\sigma^2} - y^2\frac{1}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(y\cdot\frac{\mu}{\sigma^2} - y^2\frac{1}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \ln\sigma\right) \\
b(y) &= \frac{1}{\sqrt{2\pi}} \\
\boldsymbol{\eta} &= \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} \\
T(y) &= \begin{bmatrix} x \\ x^2 \end{bmatrix} \\
a(\boldsymbol{\eta}) &= \frac{\mu^2}{2\sigma^2} + \ln\sigma = -\frac{\boldsymbol{\eta}_{(1)}^2}{4\boldsymbol{\eta}_{(2)}} + \frac{1}{2}\ln\left|\frac{1}{2\boldsymbol{\eta}_{(2)}}\right| \\
p(y;\boldsymbol{\eta}) &= \frac{1}{\sqrt{2\pi}}\exp\left(\boldsymbol{\eta}^T\begin{bmatrix} y \\ y^2 \end{bmatrix} + \frac{\boldsymbol{\eta}_{(1)}^2}{4\boldsymbol{\eta}_{(2)}} - \frac{1}{2}\ln\left|\frac{1}{2\boldsymbol{\eta}_{(2)}}\right|\right)
\end{aligned}
$$

1.3.2 Generalized Linear Models

We make following assumptions to define GLMs:

1. $p(\boldsymbol{y}_{(i)}|X_{(i)};\boldsymbol{\theta}) = \text{Exp}(\boldsymbol{\eta})$
2. $h(X_{(i)}) = \mathbb{E}[T(\boldsymbol{y}_{(i)})|X_{(i)}]$
3. $\eta = \boldsymbol{\theta}^T X_{(i)}$ or $\boldsymbol{\eta}_{(j)} = {\Theta_j}^T X_{(i)}$

Using the parameters for Bernoulli distribution, we can see that:

$$
\begin{aligned}
h(X_{(i)}) &= p(\boldsymbol{y}_{(i)} = 1|X_{(i)}) \\
&= \text{sigmoid}(\eta) \\
&= \frac{1}{1+\exp(\boldsymbol{\theta}^T X_{(i)})}
\end{aligned}
$$

and for the parameters of Gaussian distribution, the hypothesis is:

$$
\begin{aligned}
h(X_{(i)}) &= \mathbb{E}[T(\boldsymbol{y}_{(i)})|X_{(i)}] \\
&= \mu \\
&= {\Theta_1}^T X_{(i)}
\end{aligned}
$$

1.3.2.1 Multinomial Distribution and Softmax Regression Similar to Bernoulli distribution, multinomial distribution is used to represent the discrete probability distribution of a random variable $y \in 0, \ldots, k$:

$$\begin{aligned} P(y=i) &= \begin{cases} \phi_i, & i > 0 \\ 1 - \sum_{i=1}^{k} \phi_i, & i = 0 \end{cases} \\ &= \left(1 - \sum_{i=1}^{k} \phi_i\right)^{1\{y=0\}} \prod_{i=1}^{k} \phi_i^{1\{y=i\}} \end{aligned}$$

Given an identity matrix $I \in \mathcal{R}^{k \times k}$, we define $T(y)$ as:

$$\begin{aligned} T(y) &= \begin{cases} I_y, & y > 0 \\ \mathbf{0}, & y = 0 \end{cases} \\ P(y=i) &= \left(1 - \sum_{i=1}^{k} \phi_i\right)^{(1 - \mathbf{1}^T T(y))} \prod_{i=1}^{k} \phi_i^{T(y)_{(i)}} \\ &= \exp\left(\left(1 - \mathbf{1}^T T(y)\right) \ln\left(1 - \sum_{i=1}^{k} \phi_i\right) + \sum_{i=1}^{k} T(y)_{(i)} \ln \phi_i\right) \\ &= \exp\left(\ln(1 - \mathbf{1}^T \boldsymbol{\phi}) - (\ln(1 - \mathbf{1}^T \boldsymbol{\phi}) \cdot \mathbf{1})^T T(y) + T(y)^T \ln \boldsymbol{\phi}\right) \\ &= \exp\left(T(y)^T \left(\ln \frac{\boldsymbol{\phi}}{1 - \mathbf{1}^T \boldsymbol{\phi}}\right) - (-\ln(1 - \mathbf{1}^T \boldsymbol{\phi}))\right) \end{aligned}$$

so, if we define ϕ_0 as $\phi_0 = 1 - \mathbf{1}^T \boldsymbol{\phi}$:

$$\begin{aligned} b(y) &= 1 \\ \boldsymbol{\eta} &= \ln \frac{\boldsymbol{\phi}}{\phi_0} \\ a(\boldsymbol{\eta}) &= -\ln(\phi_0) \end{aligned}$$

Using the fact that $\phi_0 + \mathbf{1}^T \phi = 1$, we can derive the response function:

$$\begin{aligned} \exp(\boldsymbol{\eta}) &= \frac{\boldsymbol{\phi}}{\phi_0} \\ \mathbf{1}^T \exp(\boldsymbol{\eta}) &= \frac{\mathbf{1}^T \boldsymbol{\phi}}{\phi_0} = \frac{1 - \phi_0}{\phi_0} \\ 1 &= \phi_0 \mathbf{1}^T \exp(\boldsymbol{\eta}) + \phi_0 \\ \phi_0 &= \frac{1}{1 + \mathbf{1}^T \exp(\boldsymbol{\eta})} \\ \boldsymbol{\phi} &= \frac{\exp(\boldsymbol{\eta})}{1 + \mathbf{1}^T \exp(\boldsymbol{\eta})} \\ a(\boldsymbol{\eta}) &= \ln(1 + \mathbf{1}^T \exp(\boldsymbol{\eta})) \\ P(y; \boldsymbol{\eta}) &= \exp\left(\boldsymbol{\eta}^T T(y) - \ln(1 + \mathbf{1}^T \exp(\boldsymbol{\eta}))\right) \end{aligned}$$

Now we have paramterized it using $\boldsymbol{\eta}$, we can find its expected value, which leads to the hypothesis for the corresponding GLM:

$$\begin{aligned} h(X_{(i)}) &= \mathbb{E}[T(\boldsymbol{y}_{(i)}) | X_{(i)}; \Theta] \\ &= \sum_{\boldsymbol{y}_{(i)}=0}^{k} \frac{T(\boldsymbol{y}_{(i)}) \exp(\boldsymbol{\eta}^T T(\boldsymbol{y}_{(i)}))}{1 + \mathbf{1}^T \exp(\boldsymbol{\eta})} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{1+\mathbf{1}^T \exp(\boldsymbol{\eta})} \sum_{\boldsymbol{y}_{(i)}=1}^{k} T(\boldsymbol{y}_{(i)}) \exp(\boldsymbol{\eta}^T T(\boldsymbol{y}_{(i)})) \\
&= \frac{1}{1+\mathbf{1}^T \exp(\boldsymbol{\eta})} \sum_{\boldsymbol{y}_{(i)}=1}^{k} T(\boldsymbol{y}_{(i)}) \exp\left(\boldsymbol{\eta}_{(\boldsymbol{y}_{(i)})}\right) \\
&= \frac{\exp(\boldsymbol{\eta})}{1+\mathbf{1}^T \exp(\boldsymbol{\eta})} = \boldsymbol{\phi} \\
&= \frac{\exp(\Theta^T X_{(i)})}{1+\mathbf{1}^T \exp(\Theta^T X_{(i)})} \\
&= \begin{bmatrix}
\frac{\exp({\Theta_1}^T X_{(i)})}{1+\sum_{j=1}^{k} \exp({\Theta_j}^T X_{(i)})} \\
\frac{\exp({\Theta_2}^T X_{(i)})}{1+\sum_{j=1}^{k} \exp({\Theta_j}^T X_{(i)})} \\
\vdots \\
\frac{\exp({\Theta_k}^T X_{(i)})}{1+\sum_{j=1}^{k} \exp({\Theta_j}^T X_{(i)})}
\end{bmatrix}
\end{aligned}
$$

This model is called **Softmax Regression**, a generalization of logistic regression for multinomial distribution. Note that its parameter Θ is a $n \times k$ matrix.

To fit the parameters, we need to find the maximum likelihood estimate:

$$
\begin{aligned}
&\max_{\Theta} \prod_{i=1}^{m} P(\boldsymbol{y}_{(i)} | X_{(i)}; \Theta) \\
&= \max_{\Theta} \sum_{i=1}^{m} ((\Theta^T X_{(i)})^T T(\boldsymbol{y}_{(i)}) - \ln(1+\mathbf{1}^T \exp(\Theta^T X_{(i)}))) = \max_{\Theta} -l(\Theta) \\
\nabla_\Theta \, l(\Theta) &= -\sum_{i=1}^{m} \nabla_\Theta \, \left(\text{tr} \, \left(X_{(i)}^T \Theta \, T(\boldsymbol{y}_{(i)}) \right) - \ln(1+\mathbf{1}^T \exp(\Theta^T X_{(i)})) \right) \\
&= -\sum_{i=1}^{m} \left(X_{(i)} T(\boldsymbol{y}_{(i)})^T - \frac{X_{(i)} \exp(\Theta^T X_{(i)})^T}{1+\mathbf{1}^T \exp(\Theta^T X_{(i)})} \right) \\
&= \sum_{i=1}^{m} X_{(i)} \left(\frac{\exp(\Theta^T X_{(i)})}{1+\mathbf{1}^T \exp(\Theta^T X_{(i)})} - T(\boldsymbol{y}_{(i)}) \right)^T
\end{aligned}
$$

Similar to logistic regression, there is no general closed form solution for softmax regression. Use either gradient descent or Newton's Method to maximize the joint probability.

2 Generative Learning Algorithms

All models/algorithms mentioned before in the last chapter are **discriminative learning algorithms**, which try to maximize the joint probability of $P(\boldsymbol{y}|X;\eta)$. This chapter will introduce algorithms are **generative learning algorithms** that models $P(X|\boldsymbol{y})$ and $P(\boldsymbol{y})$ (called **prior**), then use **Bayes rule**:

$$
P(y|x) = \frac{P(x|y)P(y)}{P(x)}
$$

to derive the the distribution of $\boldsymbol{y}$ given X. Here, $P(x)$ can be calculated by applying law of total probability:

$$
P(x) = \int_y y P(x|y)
$$

2.1 Multivariate Normal Distribution

A multivariate distribution a vector of random variables outputing values in an n-dimensional space. To describe variance of high-dimensional data, we define **covariance** as the average outer product of its deviation (from its average) with itself, that is:

$$\text{Cov}(\boldsymbol{Z}) = \mathbb{E}[(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])(\boldsymbol{Z} - \mathbb{E}[\boldsymbol{Z}])^T]$$

We can derive another form of the covariance:

$$\begin{aligned}
\text{Cov}(\boldsymbol{Z}) &= \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T - \mathbb{E}[\boldsymbol{Z}]\boldsymbol{Z}^T - \boldsymbol{Z}\mathbb{E}[\boldsymbol{Z}]^T + \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T] \\
&= \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] - \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T - \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T + \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T \\
&= \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] - \mathbb{E}[\boldsymbol{Z}]\mathbb{E}[\boldsymbol{Z}]^T
\end{aligned}$$

Multivariate normal distribution is defined as following:

$$\begin{aligned}
P(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) &= \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right)
\end{aligned}$$

Multivariate normal distribution is an analogy of the single-variable counterpart in higher dimensional space. The probability density of $\boldsymbol{y}$ is highest at mean vector $\boldsymbol{\mu}$, decreases along a bell curve as $\boldsymbol{y}$ moves away from the mean vector. We see the probability density of a value $\boldsymbol{y}$ that is d units away from $\boldsymbol{\mu}$ along the unit vector $\hat{\boldsymbol{u}}$ is:

$$\begin{aligned}
P(\boldsymbol{y} = d\hat{\boldsymbol{u}} + \boldsymbol{\mu}; \boldsymbol{\mu}, \Sigma) &= \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(d\hat{\boldsymbol{u}})^T\Sigma^{-1}(d\hat{\boldsymbol{u}})\right) \\
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{d^2}{2}\hat{\boldsymbol{u}}^T\Sigma^{-1}\hat{\boldsymbol{u}}\right)
\end{aligned}$$

Since Σ is symmetric and postive semi-definite by definition, there is $\Sigma = Q\Lambda Q^T$, where the column vectors of Q are the eigenvectors of Σ and Λ is a diagonal matrix formed by the respective eigenvalues, all of which are greater or equal to 0:

$$\begin{aligned}
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{d^2}{2}\hat{\boldsymbol{u}}^TQ\Lambda^{-1}Q^T\hat{\boldsymbol{u}}\right) \\
&= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{d^2}{2}\hat{\boldsymbol{u}}^TQ\Lambda^{-1/2}\Lambda^{-1/2}Q^T\hat{\boldsymbol{u}}\right)
\end{aligned}$$

Note that since $\Lambda = \text{Diag}(\boldsymbol{\lambda})$, $\Lambda^{-1} = \text{Diag}(1/\boldsymbol{\lambda})$ and $\Lambda^{-1/2} = \text{Diag}(\boldsymbol{\lambda}^{-1/2})$, so $\Lambda^{-1/2}$ is also a symmetric, positive semi-definite matrix:

$$=$$

2.1.1 Gaussian Discriminant Analysis

$$\begin{aligned}
y &\sim \text{Bernoulli}(\phi) \\
\boldsymbol{x}|y = 0 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma) \\
\boldsymbol{x}|y = 1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma)
\end{aligned}$$

(Negative) Joint Log Probability:

$$\begin{aligned}
l(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) &= -\ln\prod_{i=1}^{m} p(X_{(i)}|\boldsymbol{y}_{(i)}; \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma)p(\boldsymbol{y_{(i)}}; \phi) \\
&= -\ln\prod_{i=1}^{m} p(X_{(i)}|\boldsymbol{y}_{(i)} = 1; \boldsymbol{\mu}_1, \Sigma)^{\boldsymbol{y}_{(i)}}p(X_{(i)}|\boldsymbol{y}_{(i)} = 0; \boldsymbol{\mu}_1, \Sigma)^{(1\,-\,\boldsymbol{y}_{(i)})}p(\boldsymbol{y_{(i)}}; \phi) \\
&= -\sum_{i=1}^{m}\left(-\frac{n}{2}\boldsymbol{y}_{(i)}\ln 2\pi - \frac{1}{2}\boldsymbol{y}_{(i)}\ln|\Sigma| - \boldsymbol{y}_{(i)}\left(\frac{1}{2}(X_{(i)} - \boldsymbol{\mu}_1)^T\Sigma^{-1}(X_{(i)} - \boldsymbol{\mu}_1)\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&- \sum_{i=1}^{m} \left(-\frac{n}{2}(1-\boldsymbol{y}_{(i)}) \ln 2\pi - \frac{1}{2}(1-\boldsymbol{y}_{(i)}) \ln|\Sigma| - (1-\boldsymbol{y}_{(i)}) \left(\frac{1}{2}(X_{(i)}-\boldsymbol{\mu}_0)^T \Sigma^{-1}(X_{(i)}-\boldsymbol{\mu}_0) \right) \right) \\
&- \sum_{i=1}^{m} \Big(\boldsymbol{y}_{(i)} \ln\phi + (1-\boldsymbol{y}_{(i)}) \ln(1-\phi) \Big) \\
&= \sum_{i=1}^{m} \left(\frac{n}{2} \ln 2\pi + \frac{1}{2} \ln|\Sigma| + \boldsymbol{y}_{(i)} \left(\frac{1}{2}(X_{(i)}-\boldsymbol{\mu}_1)^T \Sigma^{-1}(X_{(i)}-\boldsymbol{\mu}_1) \right) \right. \\
&\qquad\qquad + (1-\boldsymbol{y}_{(i)}) \left(\frac{1}{2}(X_{(i)}-\boldsymbol{\mu}_0)^T \Sigma^{-1}(X_{(i)}-\boldsymbol{\mu}_0) \right) \\
&\qquad\qquad \left. - \boldsymbol{y}_{(i)} \ln\phi - (1-\boldsymbol{y}_{(i)}) \ln(1-\phi) \right) \\
&= \frac{mn}{2} \ln 2\pi + \frac{m}{2} \ln|\Sigma| + \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{y}_{(i)} (X - \mathbf{1}\boldsymbol{\mu}_1^T)_{(i)}{}^T \Sigma^{-1} (X - \mathbf{1}\boldsymbol{\mu}_1^T)_{(i)} \\
&\qquad\qquad + \frac{1}{2} \sum_{i=1}^{m} (1-\boldsymbol{y})_{(i)} (X - \mathbf{1}\boldsymbol{\mu}_0^T)_{(i)}{}^T \Sigma^{-1} (X - \mathbf{1}\boldsymbol{\mu}_0^T)_{(i)} \\
&\qquad\qquad - \mathbf{1}^T (\boldsymbol{y} \ln\phi + (1-\boldsymbol{y}) \ln(1-\phi)) \\
&= \frac{mn}{2} \ln 2\pi + \frac{m}{2} \ln|\Sigma| + \frac{1}{2} \sum_{i=1}^{m} ((X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \mathrm{Diag}(\boldsymbol{y}))_i{}^T (\Sigma^{-1}(X - \mathbf{1}\boldsymbol{\mu}_1^T)^T)_i \\
&\qquad\qquad + \frac{1}{2} \sum_{i=1}^{m} ((X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \mathrm{Diag}(1-\boldsymbol{y}))_i{}^T (\Sigma^{-1}(X - \mathbf{1}\boldsymbol{\mu}_0^T)^T)_i \\
&\qquad\qquad - \mathbf{1}^T (\boldsymbol{y} \ln\phi + (1-\boldsymbol{y}) \ln(1-\phi)) \\
&= \frac{mn}{2} \ln 2\pi + \frac{m}{2} \ln|\Sigma| + \frac{1}{2} \mathrm{tr}\ \mathrm{Diag}(\boldsymbol{y}) (X - \mathbf{1}\boldsymbol{\mu}_1^T) \Sigma^{-1} (X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \\
&\qquad\qquad + \frac{1}{2} \mathrm{tr}\ \mathrm{Diag}(1-\boldsymbol{y}) (X - \mathbf{1}\boldsymbol{\mu}_0^T) \Sigma^{-1} (X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \\
&\qquad\qquad - \mathbf{1}^T (\boldsymbol{y} \ln\phi + (1-\boldsymbol{y}) \ln(1-\phi))
\end{aligned}
$$

Gradient with respect to parameters:

$$
\begin{aligned}
\mathrm{d}l(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) &= -\mathbf{1}^T \left(\frac{\boldsymbol{y}}{\phi} - \frac{1-\boldsymbol{y}}{1-\phi} \right) \mathrm{d}\phi \\
\mathbf{1}^T \boldsymbol{y}(1-\phi) &= \mathbf{1}^T (1-\boldsymbol{y})\phi \\
\mathbf{1}^T (2\boldsymbol{y} - 1)\phi &= \mathbf{1}^T \boldsymbol{y} \\
\mathbf{1}^T \boldsymbol{y} - \mathbf{1}^T \boldsymbol{y}\phi &= \mathbf{1}^T \mathbf{1}\phi - \mathbf{1}^T \boldsymbol{y}\phi \\
\phi &= \frac{\mathbf{1}^T \boldsymbol{y}}{\mathbf{1}^T \mathbf{1}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{\boldsymbol{y}_{(i)} = 1\}
\end{aligned}
$$

Gradient with respect to $\boldsymbol{\mu}_1$

$$
\begin{aligned}
\mathrm{d}l(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) &= \frac{1}{2} \mathrm{d}\ \mathrm{tr}\ \mathrm{Diag}(\boldsymbol{y}) (X - \mathbf{1}\boldsymbol{\mu}_1^T) \Sigma^{-1} (X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \\
&= \frac{1}{2} \mathrm{tr}\ \mathrm{d}\Big(\mathrm{Diag}(\boldsymbol{y}) \mathbf{1}\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 \mathbf{1}^T - \mathrm{Diag}(\boldsymbol{y}) X \Sigma^{-1} \boldsymbol{\mu}_1 \mathbf{1}^T - \mathrm{Diag}(\boldsymbol{y}) \mathbf{1}\boldsymbol{\mu}_1^T \Sigma^{-1} X^T \Big) \\
&= \frac{1}{2} \mathrm{d}\Big(\mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) \mathbf{1}\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 - \mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) X \Sigma^{-1} \boldsymbol{\mu}_1 - \mathrm{tr}\ \boldsymbol{\mu}_1^T \Sigma^{-1} X^T \mathrm{Diag}(\boldsymbol{y}) \mathbf{1} \Big) \\
&= \frac{1}{2} \Big(2\ \mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) \mathbf{1}\boldsymbol{\mu}_1^T \Sigma^{-1}\ \mathrm{d}\boldsymbol{\mu}_1 - \mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) X \Sigma^{-1}\ \mathrm{d}\boldsymbol{\mu}_1 - \mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) X \Sigma^{-1}\ \mathrm{d}\boldsymbol{\mu}_1 \Big) \\
&= \mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) \mathbf{1}\boldsymbol{\mu}_1^T \Sigma^{-1}\ \mathrm{d}\boldsymbol{\mu}_1 - \mathrm{tr}\ \mathbf{1}^T \mathrm{Diag}(\boldsymbol{y}) X \Sigma^{-1}\ \mathrm{d}\boldsymbol{\mu}_1 \\
&= \mathrm{tr}\ \boldsymbol{y}^T (\mathbf{1}\boldsymbol{\mu}_1^T - X) \Sigma^{-1}\ \mathrm{d}\boldsymbol{\mu}_1 \\
\nabla_{\boldsymbol{\mu}_1} l(\boldsymbol{\mu}_1) &= \Sigma^{-1} (\mathbf{1}\boldsymbol{\mu}_1^T - X)^T \boldsymbol{y}
\end{aligned}
$$

Since Σ is not singular, (or we assume so), $(\mathbf{1}\boldsymbol{\mu}_1^T - X)^T \boldsymbol{y}$ must be 0 for the product to be 0:

$$
(\mathbf{1}\boldsymbol{\mu}_1^T - X)^T \boldsymbol{y} = 0
$$

$$\boldsymbol{\mu}_1 \mathbf{1}^T \boldsymbol{y} = X^T \boldsymbol{y}$$
$$\boldsymbol{\mu}_1 = \frac{X^T \boldsymbol{y}}{\mathbf{1}^T \boldsymbol{y}} = \frac{\sum_{i=1}^m 1\{\boldsymbol{y}_{(i)} = 1\} X_{(i)}}{\sum_{i=1}^m 1\{\boldsymbol{y}_{(i)} = 1\}}$$

Via a similar process, the gradient with respect to $\boldsymbol{\mu}_0$ is:

$$\begin{aligned}
\mathrm{d}l(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) &= \operatorname{tr}\ (1-\boldsymbol{y})^T (\mathbf{1}\boldsymbol{\mu}_0^T - X)\Sigma^{-1}\,\mathrm{d}\boldsymbol{\mu}_0 \\
\nabla_{\boldsymbol{\mu}_0} l(\boldsymbol{\mu}_0) &= \Sigma^{-1}(\mathbf{1}\boldsymbol{\mu}_0^T - X)^T (1-\boldsymbol{y}) \\
0 &= (\mathbf{1}\boldsymbol{\mu}_1^T - X)^T (1-\boldsymbol{y}) \\
\boldsymbol{\mu_0} &= \frac{X^T(1-\boldsymbol{y})}{\mathbf{1}^T(1-\boldsymbol{y})} = \frac{\sum_{i=1}^m 1\{\boldsymbol{y}_{(i)} = 0\} X_{(i)}}{\sum_{i=1}^m 1\{\boldsymbol{y}_{(i)} = 0\}}
\end{aligned}$$

Gradient with respect to Σ:

$$\begin{aligned}
\mathrm{d}l(\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) &= \frac{m}{2}\operatorname{tr}\ \Sigma^{-1}\,\mathrm{d}\Sigma - \frac{1}{2}\operatorname{tr}\ \operatorname{Diag}(\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_1^T)\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}(X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \\
&\quad - \frac{1}{2}\operatorname{tr}\ \operatorname{Diag}(1-\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_0^T)\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}(X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \\
&= \frac{m}{2}\operatorname{tr}\ \Sigma^{-1}\Sigma\Sigma^{-1}\,\mathrm{d}\Sigma - \frac{1}{2}\operatorname{tr}\ \Sigma^{-1}(X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \operatorname{Diag}(\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_1^T)\Sigma^{-1}\,\mathrm{d}\Sigma \\
&\quad - \frac{1}{2}\operatorname{tr}\ \Sigma^{-1}(X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \operatorname{Diag}(1-\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_0^T)\Sigma^{-1}\,\mathrm{d}\Sigma \\
&= \operatorname{tr}\ \Sigma^{-1}\left(m\Sigma - (X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \operatorname{Diag}(\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_1^T) - (X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \operatorname{Diag}(1-\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_0^T)\right)\Sigma^{-1}\,\mathrm{d}\Sigma \\
\nabla_\Sigma l(\Sigma) &= \Sigma^{-1}\left(m\Sigma - (X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \operatorname{Diag}(\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_1^T) - (X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \operatorname{Diag}(1-\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_0^T)\right)\Sigma^{-1} \\
\Sigma &= \frac{1}{m}\left((X - \mathbf{1}\boldsymbol{\mu}_1^T)^T \operatorname{Diag}(\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_1^T) + (X - \mathbf{1}\boldsymbol{\mu}_0^T)^T \operatorname{Diag}(1-\boldsymbol{y})(X - \mathbf{1}\boldsymbol{\mu}_0^T)\right) \\
&= \frac{1}{m}(X - (\boldsymbol{y}\boldsymbol{\mu}_1^T + (1-\boldsymbol{y})\boldsymbol{\mu}_0^T))(X - (\boldsymbol{y}\boldsymbol{\mu}_1^T + (1-\boldsymbol{y})\boldsymbol{\mu}_0^T))^T
\end{aligned}$$

Gaussian discriminant analysis is a special case of logistic regression. It assumes a stronger assumption that the data is normally-distributed around the respective center of mass of the classes. When this assumption is right, the GDA is **asymptotically efficient**, which means there is no better algorithms that GDA in terms of accuracy given a set amount of training data.
However, when this assumption is not true, logistic regression finds its better fit, making it more robust when the distribution of the data is not well known.

2.2 Naive Bayes

$$\begin{aligned}
y &\sim \text{Bernoulli}(\phi) \\
\boldsymbol{x}_{(j)}|y = 0 &\sim \text{Bernoulli}(\phi_{j|y=0}) \\
\boldsymbol{x}_{(j)}|y = 1 &\sim \text{Bernoulli}(\phi_{j|y=1})
\end{aligned}$$

Naive Bayes Assumption: $\boldsymbol{x}_{(j)}$'s are conditionally independent given y, i.e.

$$p(\boldsymbol{x}|y) = \prod_{j=1}^n p(\boldsymbol{x}_{(j)}|y)$$

Although Naive Bayes Assumption is almost never strictly true, the algorithm resulting from usually works well on many problems. Now let's maximize the joint probability of the model and gradients for our parameters:

$$\begin{aligned}
&\min l(\phi, {\boldsymbol{\phi}_0}^T = [\phi_{1|y=0} \dots \phi_{n|y=0}], {\boldsymbol{\phi}_1}^T = [\phi_{1|y=1} \dots \phi_{n|y=1}]) \\
&= -\ln \prod_{i=1}^m p(X_{(i)}|\boldsymbol{y}_{(i)}) p(\boldsymbol{y}_{(i)}) \\
&= -\ln \prod_{i=1}^m \prod_{j=1}^n p(X_{ij}|\boldsymbol{y}_{(i)}) \phi^{\boldsymbol{y}_{(i)}} (1-\phi)^{1-\boldsymbol{y}_{(i)}}
\end{aligned}$$

$$
\begin{aligned}
&= -\ln \prod_{i=1}^{m}\prod_{j=1}^{n} p(X_{ij}|\boldsymbol{y}_{(i)}=1)^{\boldsymbol{y}_{(i)}} p(X_{ij}|\boldsymbol{y}_{(i)}=0)^{1-\boldsymbol{y}_{(i)}} \phi^{\boldsymbol{y}_{(i)}}(1-\phi)^{1-\boldsymbol{y}_{(i)}} \\
&= -\ln \prod_{i=1}^{m}\prod_{j=1}^{n} {\phi_{j|y=1}}^{\boldsymbol{y}_{(i)}X_{ij}}(1-\phi_{j|y=1})^{\boldsymbol{y}_{(i)}(1-X_{ij})}{\phi_{j|y=0}}^{(1-\boldsymbol{y}_{(i)})X_{ij}}(1-\phi_{j|y=0})^{(1-\boldsymbol{y}_{(i)})(1-X_{ij})}\phi^{\boldsymbol{y}_{(i)}}(1-\phi)^{1-\boldsymbol{y}_{(i)}} \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{n}\Big(\boldsymbol{y}_{(i)}X_{ij}\ln\phi_{j|y=1} + \boldsymbol{y}_{(i)}(1-X_{ij})\ln(1-\phi_{j|y=1}) + (1-\boldsymbol{y}_{(i)})X_{ij}\ln\phi_{j|y=0} + (1-\boldsymbol{y}_{(i)})(1-X_{ij})\ln(1-\phi_{j|y=0})\Big) \\
&\quad - n\sum_{i=1}^{m}\Big(\boldsymbol{y}_{(i)}\ln\phi + (1-\boldsymbol{y}_{(i)})\ln(1-\phi)\Big) \\
&= -\sum_{i=1}^{m}\Big(\boldsymbol{y}_{(i)}{X_{(i)}}^{T}\ln\boldsymbol{\phi}_1 + \boldsymbol{y}_{(i)}(1-X_{(i)})^{T}\ln(1-\boldsymbol{\phi}_1) + (1-\boldsymbol{y}_{(i)}){X_{(i)}}^{T}\ln\boldsymbol{\phi}_0 + (1-\boldsymbol{y}_{(i)})(1-X_{(i)})^{T}\ln(1-\boldsymbol{\phi}_0)\Big) \\
&\quad - n(\mathbf{1}^T\boldsymbol{y}\ln\phi + \mathbf{1}^T(1-\boldsymbol{y})\ln(1-\phi)) \\
&= -\boldsymbol{y}^T(X\ln\boldsymbol{\phi}_1 - (1-X)\ln(1-\boldsymbol{\phi}_1)) - (1-\boldsymbol{y})^T(X\ln\boldsymbol{\phi}_0 + (1-X)\ln(1-\boldsymbol{\phi}_0)) - n(\mathbf{1}^T\boldsymbol{y}\ln\phi + \mathbf{1}^T(1-\boldsymbol{y})\ln(1-\phi))
\end{aligned}
$$

Gradient:

$$
\begin{aligned}
&\mathrm{d}l(\phi,\boldsymbol{\phi}_0,\boldsymbol{\phi}_1) \\
&= \boldsymbol{y}^T\left(X\,\mathrm{Diag}\left(\frac{1}{\boldsymbol{\phi}_1}\right)\mathrm{d}\boldsymbol{\phi}_1 - (1-X)\,\mathrm{Diag}\left(\frac{1}{1-\boldsymbol{\phi}_1}\right)\mathrm{d}\boldsymbol{\phi}_1\right) + (1-\boldsymbol{y}^T)\left(X\,\mathrm{Diag}\left(\frac{1}{\boldsymbol{\phi}_0}\right)\mathrm{d}\boldsymbol{\phi}_0 - (1-X)\,\mathrm{Diag}\left(\frac{1}{1-\boldsymbol{\phi}_0}\right)\mathrm{d}\boldsymbol{\phi}_0\right) \\
&\quad + n\left(\mathbf{1}^T\boldsymbol{y}\frac{\mathrm{d}\phi}{\phi} - \mathbf{1}^T(1-\boldsymbol{y})\frac{\mathrm{d}\phi}{1-\phi}\right)
\end{aligned}
$$

Solution for $\boldsymbol{\phi}_1$:

$$
\begin{aligned}
\mathrm{Diag}\left(\frac{1}{\boldsymbol{\phi}_1}\right)X^T\boldsymbol{y} &= \mathrm{Diag}\left(\frac{1}{1-\boldsymbol{\phi}_1}\right)(1-X)^T\boldsymbol{y} \\
\mathrm{Diag}\,(1-\boldsymbol{\phi}_1)\,X^T\boldsymbol{y} &= \mathrm{Diag}\,(\boldsymbol{\phi}_1)\,(1-X)^T\boldsymbol{y} \\
X^T\boldsymbol{y} &= \mathrm{Diag}\,(\boldsymbol{\phi}_1)\,((1-X)^T\boldsymbol{y} + X^T\boldsymbol{y}) \\
X^T\boldsymbol{y} &= \boldsymbol{\phi}_1 \circ ({1_X}^T\boldsymbol{y}) = \boldsymbol{\phi}_1(1^T\boldsymbol{y}) \\
\boldsymbol{\phi}_1 &= \frac{X^T\boldsymbol{y}}{\mathbf{1}^T\boldsymbol{y}}
\end{aligned}
$$

Similarly:

$$
\boldsymbol{\phi}_2 = \frac{X^T(1-\boldsymbol{y})}{\mathbf{1}^T(1-\boldsymbol{y})}
$$

For ϕ, we have:

$$
\begin{aligned}
\frac{\mathbf{1}^T\boldsymbol{y}}{\phi} &= \frac{\mathbf{1}^T(1-\boldsymbol{y})}{1-\phi} \\
\mathbf{1}^T\boldsymbol{y} - \mathbf{1}^T\boldsymbol{y}\phi &= \mathbf{1}^T(1-\boldsymbol{y})\phi \\
\mathbf{1}^T\boldsymbol{y} &= \mathbf{1}^T\mathbf{1}_{\boldsymbol{y}}\phi = m\phi \\
\phi &= \frac{\mathbf{1}^T\boldsymbol{y}}{m}
\end{aligned}
$$

3 Support Vector Machine

3.1 Preliminaries

3.1.1 Lagrange Multiplier

Lagrange Multiplier is a method of finding critical points of a scalar-valued function subject to equality constraints, namely

$$
\max_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{given } g(\boldsymbol{x}) = 0
$$

or:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad \text{given } g(\boldsymbol{x}) = 0$$

Such critical points always lies on points where the contours of the two functions are parallel, and since the contour of function is always perpendicular to the gradient at any point, that means we need to find point $\boldsymbol{x}'$ such that:

$$\begin{aligned} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}') &= \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}') \\ g(\boldsymbol{x}') &= 0 \end{aligned}$$

This can be cleverly written in term of the gradient of **Lagrangian** being 0:

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \lambda) &= f(\boldsymbol{x}) - \lambda g(\boldsymbol{x}) \\ \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}', \lambda) &= \begin{bmatrix} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}') - \lambda \nabla_{\boldsymbol{x}} g(\boldsymbol{x}') \\ g(\boldsymbol{x}') \end{bmatrix} = 0 \end{aligned}$$

The solutions can be maxima, minima, or saddles, but if $g(\boldsymbol{x}) = 0$ contains a closed and bounded region, there has to be at least 2 solutions, corresponding to its minimum and maximum.

3.1.2 Distance of a Point to a Hyperplane

Using Lagrange Multiplier, Let's calculate the distance of a point $\boldsymbol{x}_0$ in to a hyperplane $\boldsymbol{\theta}^T \boldsymbol{x} + b = 0$

$$\min_{\boldsymbol{x}} \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 \text{ given } \boldsymbol{\theta}^T \boldsymbol{x} + b = 0$$

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \lambda) &= (\boldsymbol{x} - \boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) - \lambda(\boldsymbol{\theta}^T \boldsymbol{x} + b) \\ \nabla_{\boldsymbol{x},\lambda} \mathcal{L}(\boldsymbol{x}, \lambda) &= \begin{bmatrix} 2(\boldsymbol{x} - \boldsymbol{x}_0) - \lambda \boldsymbol{\theta} \\ -(\boldsymbol{\theta}^T \boldsymbol{x} + b) \end{bmatrix} = 0 \\ \lambda \boldsymbol{\theta} &= 2(\boldsymbol{x} - \boldsymbol{x}_0) \\ \lambda &= \frac{2\boldsymbol{\theta}^T (\boldsymbol{x} - \boldsymbol{x}_0)}{\boldsymbol{\theta}^T \boldsymbol{\theta}} = \frac{-2(b + \boldsymbol{\theta}^T \boldsymbol{x}_0)}{\boldsymbol{\theta}^T \boldsymbol{\theta}} \qquad \boldsymbol{\theta}^T \boldsymbol{x} = -b \\ \frac{\lambda \boldsymbol{\theta}}{2} &= (\boldsymbol{x} - \boldsymbol{x}_0) \\ \frac{\lambda^2 \boldsymbol{\theta}^T \boldsymbol{\theta}}{4} &= (\boldsymbol{x} - \boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) \\ (\boldsymbol{x} - \boldsymbol{x}_0)^T (\boldsymbol{x} - \boldsymbol{x}_0) &= \frac{(\boldsymbol{\theta}^T \boldsymbol{x}_0 + b)^2}{\boldsymbol{\theta}^T \boldsymbol{\theta}} \\ \|\boldsymbol{x} - \boldsymbol{x}_0\| &= \frac{|\boldsymbol{\theta}^T \boldsymbol{x}_0 + b|}{\|\boldsymbol{\theta}\|} \end{aligned}$$

3.1.3 Lagrange duality

Given an optimization problem of following form:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f(\boldsymbol{w}) \\ \text{s.t.} \quad & \boldsymbol{g}(\boldsymbol{w}) \leq 0 \\ & \boldsymbol{h}(\boldsymbol{w}) = 0 \end{aligned}$$

define its **generalized Lagrangian** to be:

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{w}) + \boldsymbol{\alpha}^T \boldsymbol{g}(\boldsymbol{w}) + \boldsymbol{\beta}^T \boldsymbol{h}(\boldsymbol{w})$$

Note that if we try to maximize the Lagrangian with the condition that $\boldsymbol{\alpha} \geq 0$, it is only bounded if the constraints are satisfied, otherwise, that it is bounded, it maximize to $f(\boldsymbol{w})$, in other words, we have:

$$\theta_{\mathcal{P}}(\boldsymbol{w}) = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \boldsymbol{\alpha} \geq 0} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} f(\boldsymbol{w}) & \boldsymbol{g}(\boldsymbol{w}) \leq 0 \wedge \boldsymbol{h}(\boldsymbol{w}) = 0 \\ \infty & \text{otherwise} \end{cases}$$

Minimizing this problem is exactly same as minimizing $f(\boldsymbol{w})$ conforming to the given constraints, and this is what we call **primal problem**:

$$\min_{\boldsymbol{w}} \theta_{\mathcal{P}}(\boldsymbol{w}) = \min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{a}\geq 0} \mathcal{L}(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})$$

If we reverse the order of min and max, we have the **dual problem**:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{a}\geq 0} \theta_{\mathcal{D}}(\boldsymbol{w}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{a}\geq 0} \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta})$$

Let $p*$ be the solution to the primal problem and $d*$ be the solution to the dual problem:

$$d^* = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{a}\geq 0} \min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}) \leq \min_{\boldsymbol{w}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{a}\geq 0} \mathcal{L}(\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\beta}) = p^*$$

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}:\boldsymbol{a}\geq 0} \min_{\boldsymbol{w}} \Big(f(\boldsymbol{w}) + \boldsymbol{\alpha}^T \boldsymbol{g}(\boldsymbol{w}) + \boldsymbol{\beta}^T \boldsymbol{h}(\boldsymbol{w}) \Big)$$

This property is called weak duality, which holds for any continuous function $\mathcal{L}$. However when **Slater's condition** are met:

$$\begin{aligned} &f \text{ and } \boldsymbol{g}_{(i)} \text{ are convex for all } i \\ &\boldsymbol{h}_{(i)} \text{ is affine for all } i \\ &\text{there exists } \boldsymbol{w} \text{ such that } \boldsymbol{g}_{(i)}(\boldsymbol{w}) < 0 \text{ for all } i \end{aligned}$$

then, **strong duality** holds, i.e:

$$d* = p*$$

Another condition, that are both sufficient and necessary when strong duality hold is the **Karush-Kuhn-Tucker condtion**, formulated as following, given a feasible solution $\boldsymbol{w}^*$:

$$\begin{aligned} \nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}^*,\boldsymbol{\alpha},\boldsymbol{\beta}) &= 0 \\ \boldsymbol{\alpha} \circ \boldsymbol{g}(\boldsymbol{w}^*) &= 0 \\ \boldsymbol{\alpha} &\geq 0 \end{aligned}$$

3.2 Optimal Margin Classifier

Given a **linearly separable** dataset $(X_{(i)}, \boldsymbol{y}_{(i)}) \in (\mathbb{R}^n, \{-1, 1\})$, let $\boldsymbol{\theta}$ and b define a hyperplane that properly separates this dataset. The hypothesis for this classifier is:

$$h_{\boldsymbol{\theta}}(X_{(i)}) = \text{sign}(\boldsymbol{\theta}^T X_{(i)} + b)$$

We define the **geometric margin** of the ith data point to be:

$$\gamma_i = \frac{\boldsymbol{y}_{(i)}(\boldsymbol{\theta}^T X_{(i)} + b)}{\|\boldsymbol{\theta}\|}$$

Note that margin for a particular data point is the distance of that data point to the separating hyperplane when the hypothesis matches its label, and the negative of that distance when the hypothesis mismatches. We can use this to denotes the confidence we have towards the hypothesis for this particular data point. For the entire dataset, we use the minimal margin in the dataset:

$$\gamma = \min_{i=1,\ldots,m} \frac{\boldsymbol{y}_{(i)}(\boldsymbol{\theta}^T X_{(i)} + b)}{\|\boldsymbol{\theta}\|} = \frac{1}{\|\boldsymbol{\theta}\|} \min_{i=1,\ldots,m} \boldsymbol{y}_{(i)}(\boldsymbol{\theta}^T X_{(i)} + b) = \frac{\hat{\gamma}}{\|\boldsymbol{\theta}\|}$$

To find the separating hyperplane, we maximize the margin for our dataset:

$$\begin{aligned} &\max_{\hat{\gamma},\boldsymbol{\theta}} \frac{\hat{\gamma}}{\|\boldsymbol{\theta}\|} \\ &\text{s.t. } \boldsymbol{y}_{(i)}(\boldsymbol{\theta}^T X_{(i)} + b) \geq \hat{\gamma} \end{aligned}$$

Since our model is invariant to scaling, let's scale it down by $\hat{\gamma}$:

$$\begin{aligned} &\max_{\hat{\gamma},\boldsymbol{\theta}} \frac{1}{\|\boldsymbol{\theta}\|} \\ &\text{s.t. } \boldsymbol{y}_{(i)}(\boldsymbol{\theta}^T X_{(i)} + b) \geq 1 \end{aligned}$$

Then we change maximize $\|\boldsymbol{\theta}\|$ to minimize $\|\boldsymbol{\theta}\|^2$ to allow our model solved by quadratic programming:

$$\begin{aligned} &\min_{\hat{\gamma},\boldsymbol{\theta}} \frac{1}{2} \boldsymbol{\theta}^T \boldsymbol{\theta} \\ &\text{s.t. } X\boldsymbol{\theta} + b \leq -\boldsymbol{y} \end{aligned}$$

3.2.1 Lagrange Dual Form

Let's write our primal in a more amenable form:

$$\begin{aligned}&\min \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}\\&\text{s.t. } 1-(X\boldsymbol{\theta}+b)\circ\boldsymbol{y}\leq 0\end{aligned}$$

From the constraints of this problem, we can see that only $\boldsymbol{g}(\boldsymbol{\theta})_{(i)}=0$ only when $X_{(i)}$ is a support vector, so to satisfy $\boldsymbol{a}\circ\boldsymbol{g}(\boldsymbol{\theta})=0$, $\boldsymbol{\alpha}_{(i)}$ must be zero unless $X_{(i)}$ is a support vector.

The Lagrangian is:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\alpha})&=\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta}+\boldsymbol{\alpha}^T(1-(X\boldsymbol{\theta}+b)\circ\boldsymbol{y})\\\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\alpha})&=\boldsymbol{\theta}-X^T\operatorname{Diag}(\boldsymbol{y})\boldsymbol{\alpha}=0\\\boldsymbol{\theta}&=X^T\operatorname{Diag}(\boldsymbol{y})\boldsymbol{\alpha}\\\nabla_b\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\alpha})&=\boldsymbol{y}^T\boldsymbol{\alpha}=0\end{aligned}$$

Plugging the definition of $\boldsymbol{\theta}$ to the Lagrangian, we have:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha})&=\frac{1}{2}\boldsymbol{\alpha}^T\operatorname{Diag}(\boldsymbol{y})XX^T\operatorname{Diag}(\boldsymbol{y})\boldsymbol{\alpha}+\boldsymbol{\alpha}^T\mathbf{1}-\boldsymbol{\alpha}^T\operatorname{Diag}(\boldsymbol{y})XX^T\operatorname{Diag}(\boldsymbol{y})\boldsymbol{\alpha}-(\boldsymbol{\alpha}^T\boldsymbol{y})b\\&=\boldsymbol{\alpha}^T\mathbf{1}-\frac{1}{2}(\boldsymbol{\alpha}\circ\boldsymbol{y})^TXX^T(\boldsymbol{\alpha}\circ\boldsymbol{y})\end{aligned}$$

To find $\boldsymbol{\alpha}^*$, maximize the dual problem:

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\mathbf{1}-\frac{1}{2}(\boldsymbol{\alpha}\circ\boldsymbol{y})^TXX^T(\boldsymbol{\alpha}\circ\boldsymbol{y})\right)\\&\text{s.t. }\boldsymbol{\alpha}\geq 0\\&\quad\boldsymbol{\alpha}^T\boldsymbol{y}=0\end{aligned}$$

then calculate $\boldsymbol{\theta}^*$ with:

$$\begin{aligned}\boldsymbol{\theta}^*&=X^T(\boldsymbol{y}\circ\boldsymbol{\alpha})\\\boldsymbol{b}^*&=-\frac{\max_{i:y_{(i)}=-1}\boldsymbol{\theta}^{*T}X_{(i)}+\min_{i:y_{(i)}=1}\boldsymbol{\theta}^{*T}X_{(i)}}{2}\end{aligned}$$

To exploit the fact that $\boldsymbol{\alpha}$ is mostly zero, let X_s denotes a matrix of support vectors (in rows), $\boldsymbol{y}_s$ and $\boldsymbol{a}_s$ denote corresponding label and $\boldsymbol{\alpha}_{(i)}$, we have:

$$\begin{aligned}X\boldsymbol{\theta}&=XX^T(\boldsymbol{y}\circ\boldsymbol{\alpha})\\&=XX_s{}^T(\boldsymbol{y}_s\circ\boldsymbol{\alpha}_s)\end{aligned}$$

where $XX_s{}^T$ is called **kernel** of the SVM; as we will see later, we can change the kernel to classify linearly inseparable data.

3.2.2 Kernel

Kernels like $K(\boldsymbol{x},\boldsymbol{z})$ are functions that map its arguments to higher dimensions then calculating their inner products. For example, define kernel K as:

$$\begin{aligned}K(\boldsymbol{x},\boldsymbol{z})&=(\boldsymbol{x}^T\boldsymbol{z}+c)^2\\&=((x_1z_1)^2+(x_2z_2)^2+(x_3z_3)^2+2x_1z_1x_2z_2+2x_1z_1x_3z_3+2x_2z_2x_3z_3+2x_1z_1c+2x_2z_2c+2x_3z_3c+c^2)\\&=\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{z})\\\boldsymbol{\phi}(\boldsymbol{x})&=\begin{bmatrix}x_1^2 & x_2^2 & x_3^2 & \sqrt{2}x_1x_2 & \sqrt{2}x_1x_3 & \sqrt{2}x_2x_3 & \sqrt{2c}x_1 & \sqrt{2c}x_2 & \sqrt{2c}x_3 & c\end{bmatrix}^T\end{aligned}$$

Kernels with higher degrees like $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T\boldsymbol{z} + c)^d$ projects the features to an $\frac{(n+d)!}{n!d!}$ feature space, while keep $O(n)$ computational complexity. There is also **Gaussian kernel**, which:

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{z}) &= \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2}{2\sigma^2}\right) \\ &= 1 - \frac{\|\boldsymbol{x} - \boldsymbol{z}\|^2}{2\sigma^2} + \frac{\|\boldsymbol{x} - \boldsymbol{z}\|^4}{4\sigma^4} - \frac{\|\boldsymbol{x} - \boldsymbol{z}\|^6}{8\sigma^6} + \dots \end{aligned}$$

maps the features onto an infinite-dimensional feature space, as shown above.
Each kernel $K(\boldsymbol{x}, \boldsymbol{z})$ is also associated with a **kernel matrix** K, defined as such:

$$\begin{aligned} &K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\text{given a list of vectors } \boldsymbol{x}_1 \dots \boldsymbol{x}_m \in \mathbb{R}^n \end{aligned}$$

We can show that K is a symmetric, positive semi-definite matrix:

$$K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)^T\boldsymbol{\phi}(\boldsymbol{x}_j) = K(\boldsymbol{x}_j, \boldsymbol{x}_i) = K_{ji}$$

To prove its definiteness, let's define $\boldsymbol{y}_1 \dots \boldsymbol{y}_n \in \mathbb{R}^m$ where $\boldsymbol{y}_j$ contains the jth elements of $\boldsymbol{\phi}(\boldsymbol{x}_1) \dots \boldsymbol{\phi}(\boldsymbol{x}_m)$, for any vector $\boldsymbol{z} \in \mathbb{R}^m$:

$$\boldsymbol{z}^T K \boldsymbol{z} = \boldsymbol{z}^T \left(\sum_{i=1}^m \boldsymbol{y}_i \boldsymbol{y}_i^T\right) \boldsymbol{z} = \sum_{i=1}^m \boldsymbol{z}^T \boldsymbol{y}_i \boldsymbol{y}_i^T \boldsymbol{z} = \sum_{i=1}^m (\boldsymbol{y}_i^T \boldsymbol{z})^T (\boldsymbol{y}_i^T \boldsymbol{z}) \geq 0$$

Thus, a kernel matrix is always positive-semidefinite. Moreover, given a symmetric, positive semi-definite matrix, its corresponding kernel function is valid:

$$\begin{aligned} \boldsymbol{z}^T K \boldsymbol{z} &= \sum_{j=1}^m \left(\boldsymbol{z}_{(j)} \sum_{i=1}^m K(\boldsymbol{x}_i, \boldsymbol{x}_j) \boldsymbol{z}_{(i)}\right) \\ &= \sum_{j=1}^m \sum_{i=1}^m \boldsymbol{z}_{(j)} K(\boldsymbol{x}_i, \boldsymbol{x}_j) \boldsymbol{z}_{(i)} \\ &= \end{aligned}$$

3.2.3 Soft Margin Classifier

Not all data are linearly separable. Instead of requiring all data points to have $\boldsymbol{y} \circ (X\boldsymbol{\theta} + b) \geq 1$, we can relax this constraint by subtracting $\boldsymbol{\xi}_{(i)}$ from each term while making such relaxations incur a penalty on the objective function by adding the sum of $\boldsymbol{\xi}$:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \boldsymbol{\xi}} \ & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C(\mathbf{1}^T\boldsymbol{\xi}) \\ \text{s.t. } & \boldsymbol{y} \circ (X\boldsymbol{\theta} + b) \geq 1 - \boldsymbol{\xi} \\ & \boldsymbol{\xi} \geq 0 \end{aligned}$$

Here C controls the weight of penalty slack variable incurs. When C is large, the classifier will behave more like an optimal margin classifier, ensuring most of the data points have a margin greater the "minimum" margin of the dataset. When C is small, the classifier will optimize towards a larger margin, despite many of the data pointing probably having a margin less than the "minimum" margin.

Let's rewrite the constraints of the problem to that of standard quadratic programming:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \boldsymbol{\xi}} \ & \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C(\mathbf{1}^T\boldsymbol{\xi}) \\ \text{s.t. } & 1 - \boldsymbol{\xi} - \boldsymbol{y} \circ (X\boldsymbol{\theta} + b) \leq 0 \\ & -\boldsymbol{\xi} \leq 0 \end{aligned}$$

The dual program of this quadratic program can be derived, again, by differentiating its Lagrangian:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C(\mathbf{1}^T\boldsymbol{\xi}) + \boldsymbol{\alpha}^T(1 - \boldsymbol{\xi} - \text{Diag}(\boldsymbol{y})X\boldsymbol{\theta} - b\boldsymbol{y}) - \boldsymbol{\gamma}^T\boldsymbol{\xi}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are our Lagrange multipliers, since we have two constraints; now calcuate the gradients and set them to 0:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\alpha}) &= \boldsymbol{\theta} - X^T \operatorname{Diag}(\boldsymbol{y})\boldsymbol{\alpha} = 0 \\
\nabla_{b}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\alpha}) &= -\boldsymbol{\alpha}^T\boldsymbol{y} = 0 \\
\nabla_{\boldsymbol{\xi}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\alpha}) &= C\mathbf{1} - \boldsymbol{\alpha} - \boldsymbol{\gamma} = 0
\end{aligned}$$

Since $b\boldsymbol{\alpha}^T\boldsymbol{y} = 0$ and $C(\mathbf{1}^T\boldsymbol{\xi}) - \boldsymbol{\alpha}^T\boldsymbol{\xi} - \boldsymbol{\gamma}^T\boldsymbol{\xi} = 0$, we get the same solution for dual program objective:

$$\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\circ\boldsymbol{y})^T XX^T(\boldsymbol{\alpha}\circ\boldsymbol{y})$$

and since $\boldsymbol{\gamma} = C - \boldsymbol{\alpha} \geq 0$, we have constraints:

$$\begin{aligned}
\text{subject to } \boldsymbol{\alpha}^T\boldsymbol{y} &= 0 \\
0 \leq \boldsymbol{\alpha} &\leq C
\end{aligned}$$

The KKT dual-complementarity condition for the dual problem is:

$$\begin{aligned}
\boldsymbol{\alpha}\circ(1 - \boldsymbol{\xi} - \boldsymbol{y}(X\boldsymbol{\theta} + b)) &= 0 \\
\boldsymbol{\gamma}\circ\boldsymbol{\xi} &= 0
\end{aligned}$$

We can make find cases to test if a data points conforms to these condition:

$$\begin{aligned}
\boldsymbol{\alpha}_{(i)} = 0 &\implies 1 - \boldsymbol{\xi}_{(i)} - (\boldsymbol{\theta}^T X_{(i)} + b) \leq 0 \wedge \boldsymbol{\gamma}_{(i)} = C > 0 \\
&\implies 1 - \boldsymbol{\xi}_{(i)} - (\boldsymbol{\theta}^T X_{(i)} + b) \leq 0 \wedge \boldsymbol{\xi}_{(i)} = 0 \\
&\implies (\boldsymbol{\theta}^T X_{(i)} + b) \geq 1 \\
0 < \boldsymbol{\alpha}_{(i)} < C &\implies 1 - \boldsymbol{\xi}_{(i)} - (\boldsymbol{\theta}^T X_{(i)} + b) = 0 \wedge \boldsymbol{\gamma}_{(i)} = C - \boldsymbol{\alpha}_{(i)} > 0 \\
&\implies 1 - \boldsymbol{\xi}_{(i)} - (\boldsymbol{\theta}^T X_{(i)} + b) = 0 \wedge \boldsymbol{\xi}_{(i)} = 0 \\
&\implies \boldsymbol{\theta}^T X_{(i)} + b = 1 \\
\boldsymbol{\alpha}_{(i)} = C &\implies 1 - \boldsymbol{\xi}_{(i)} - (\boldsymbol{\theta}^T X_{(i)} + b) = 0 \wedge \boldsymbol{\gamma}_{(i)} = C - C = 0 \\
&\implies 1 - \boldsymbol{\xi}_{(i)} - (\boldsymbol{\theta}^T X_{(i)} + b) = 0 \wedge \boldsymbol{\xi}_{(i)} \geq 0 \\
&\implies \boldsymbol{\theta}^T X_{(i)} + b \leq 1
\end{aligned}$$

3.2.4 Sequential Minimal Optimization

Sequential Minimal Optimization (SMO) is an algorithm that optimizes the dual program of SVM via coordinate ascent. Coordinate ascent optimizes a problem by optimizing each dimension of the variable sequentially, while fixing other dimensions during this process. It is an efficient algorithm for convex optimization provided the problem can be solved efficiently along each dimension. In the case SVM dual program, we optimize one **pair** of dimension a time to satisfy the constraint $\boldsymbol{\alpha}^T\boldsymbol{y} = 0$. For example, if we select $\boldsymbol{\alpha}_{(p)}, \boldsymbol{\alpha}_{(q)}$ to optimize, given the constraint $\boldsymbol{\alpha}^T\boldsymbol{y} = 0$, we have:

$$\begin{aligned}
\boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)} + \boldsymbol{\alpha}_{(q)}\boldsymbol{y}_{(q)} &= -\sum_{i=1;i\notin\{p,q\}}^{m} \boldsymbol{\alpha}_{(i)}\boldsymbol{y}_{(i)} = C_1 \\
\boldsymbol{\alpha}_{(q)} &= \frac{C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}}{\boldsymbol{y}_{(q)}} = \boldsymbol{y}_{(q)}(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}) \qquad\qquad \boldsymbol{y} \in \{-1,1\}^m
\end{aligned}$$

Plugging in this formula for $\boldsymbol{\alpha}_{(q)}$ and treating other variables as constants, our dual program becomes:

$$\max_{\boldsymbol{\alpha}_{(p)}} \boldsymbol{\alpha}_{(p)} + \boldsymbol{y}_{(q)}(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}) - \sum_{j=1}^{m} M(p,j) - \sum_{j=1}^{m} M(q,j) - \sum_{i=1}^{m} M(i,p) - \sum_{i=1}^{m} M(i,q) + M(q,p) + M(p,q) + M(p,p) + M(q,q) + C_2$$

where $M(i,j) = \boldsymbol{y}_{(i)}\boldsymbol{y}_{(j)}\boldsymbol{\alpha}_{(i)}\boldsymbol{\alpha}_{(j)}K(X_{(i)}, X_{(j)})$. Since $M(i,j) = M(j,i)$, this is equivalent to following:

$$\max_{\boldsymbol{\alpha}_{(p)}} \boldsymbol{\alpha}_{(p)} + \boldsymbol{y}_{(q)}(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}) - 2\sum_{j=1}^{m} M(p,j) - 2\sum_{j=1}^{m} M(q,j) + 2M(p,q) + M(p,p) + M(q,q) + C_2$$

$$\iff \max_{\boldsymbol{\alpha}_{(p)}} \boldsymbol{\alpha}_{(p)} + \boldsymbol{y}_{(q)}(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}) - M(p,p) - 2M(p,q) - 2\sum_{j=1;j\notin\{p,q\}}^{m} M(p,j) - M(q,q) - 2\sum_{j=1;j\notin\{p,q\}}^{m} M(q,j) + C_2$$

This is a one dimensional quadratic function, just make its derivative 0 to maximize it:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_{(p)}}\left(\boldsymbol{\alpha}_{(p)} + \boldsymbol{y}_{(q)}(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}) - M(p,p) - 2M(p,q) - 2\sum_{j=1;j\notin\{p,q\}}^{m} M(p,j) - M(q,q) - 2\sum_{j=1;j\notin\{p,q\}}^{m} M(q,j) + C_2\right)$$
$$= 1 - \boldsymbol{y}_{(p)}\boldsymbol{y}_{(q)} - \frac{\mathrm{d}M(p,p)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} + 2\frac{\mathrm{d}M(p,q)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} + \frac{\mathrm{d}M(q,q)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} + 2\sum_{j=1;j\notin\{p,q\}}^{m} \frac{\mathrm{d}M(p,j)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} + 2\sum_{j=1;j\notin\{p,q\}}^{m} \frac{\mathrm{d}M(q,j)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}}$$

where we have:

$$\begin{aligned}
\frac{\mathrm{d}M(p,p)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} &= 2\boldsymbol{y}_{(p)}^2\boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(p)}) = 2\boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(p)}) \\
\frac{\mathrm{d}M(p,q)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} &= \boldsymbol{y}_{(p)}(\boldsymbol{y}_{(q)}^2(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)}))K(X_{(p)},X_{(q)}) - (\boldsymbol{y}_{(q)}\boldsymbol{y}_{(p)})\boldsymbol{y}_{(p)}\boldsymbol{y}_{(q)}\boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(q)}) \\
&= \boldsymbol{y}_{(p)}C_1K(X_{(p)},X_{(q)}) - \boldsymbol{y}_{(p)}^2\boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(q)}) - \boldsymbol{y}_{(q)}^2\boldsymbol{y}_{(p)}^2\boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(q)}) \\
&= \boldsymbol{y}_{(p)}C_1K(X_{(p)},X_{(q)}) - \boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(q)}) - \boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(q)}) \\
&= (\boldsymbol{y}_{(p)}C_1 - 2\boldsymbol{\alpha}_{(p)})K(X_{(p)},X_{(q)}) \\
\frac{\mathrm{d}M(q,q)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} &= -2\boldsymbol{y}_{(q)}(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)})K(X_{(q)},X_{(q)})\boldsymbol{y}_{(q)}\boldsymbol{y}_{(p)} = -2(C_1 - \boldsymbol{\alpha}_{(p)}\boldsymbol{y}_{(p)})K(X_{(q)},X_{(q)})\boldsymbol{y}_{(p)} \\
&= 2\boldsymbol{\alpha}_{(p)}K(X_{(q)},X_{(q)}) - 2C_1K(X_{(q)},X_{(q)})\boldsymbol{y}_{(p)} \\
&= 2K(X_{(q)},X_{(q)})(\boldsymbol{\alpha}_{(p)} - C_1\boldsymbol{y}_{(p)}) \\
\frac{\mathrm{d}M(p,j)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} &= \boldsymbol{y}_{(p)}\boldsymbol{y}_{(j)}\boldsymbol{\alpha}_{(j)}K(X_{(p)},X_{(j)}) \\
\frac{\mathrm{d}M(q,j)}{\mathrm{d}\boldsymbol{\alpha}_{(p)}} &= -\boldsymbol{y}_{(q)}\boldsymbol{y}_{(j)}\boldsymbol{\alpha}_{(j)}K(X_{(q)},X_{(j)})
\end{aligned}$$

Plugging these derivatives in, we have:

$$\begin{aligned}
&= 1 - \boldsymbol{y}_{(p)}\boldsymbol{y}_{(q)} - 2\boldsymbol{\alpha}_{(p)}K(X_{(p)},X_{(p)}) - 2(\boldsymbol{y}_{(p)}C_1 - 2\boldsymbol{\alpha}_{(p)})K(X_{(p)},X_{(q)}) - 2K(X_{(q)},X_{(q)})(\boldsymbol{\alpha}_{(p)} - C_1\boldsymbol{y}_{(p)}) \\
&-2\sum_{j=1;j\notin\{p,q\}}^{m} \boldsymbol{y}_{(p)}\boldsymbol{y}_{(j)}\boldsymbol{\alpha}_{(j)}K(X_{(p)},X_{(j)}) - 2\sum_{j=1;j\notin\{p,q\}}^{m} \boldsymbol{y}_{(q)}\boldsymbol{y}_{(j)}\boldsymbol{\alpha}_{(j)}K(X_{(q)},X_{(j)}) = 0 \\
&(K(X_{(p)},X_{(q)}) - 2K(X_{(p)},X_{(p)}) - K(X_{(q)},X_{(q)}))\boldsymbol{\alpha}_{(p)} \\
&= \sum_{j=1;j\notin\{p,q\}}^{m} \boldsymbol{y}_{(j)}\boldsymbol{\alpha}_{(j)}(\boldsymbol{y}_{(p)}K(X_{(p)},X_{(j)}) - \boldsymbol{y}_{(q)}K(X_{(q)},X_{(j)})) - \frac{1 - \boldsymbol{y}_{(p)}\boldsymbol{y}_{(q)}}{2} + \boldsymbol{y}_{(p)}C_1K(X_{(p)},X_{(q)}) - 2K(X_{(q)},X_{(q)})C_1\boldsymbol{y}_{(p)}
\end{aligned}$$

3.2.5 Multi-class SVM (Crammer and Singer)

Let's start with an linear multi-class classifier for a dataset of $S = \{(X_{(1)}, \boldsymbol{y}_{(1)}), \ldots, (X_{(m)}, \boldsymbol{y}_{(m)})\}$ where $X \in \mathbb{R}^{m\times n}$ and $y \in \{1, \ldots, k\}^m$:

$$h_M(\boldsymbol{x}) = \arg\max_{r=1}^{k} {M_r}^T\boldsymbol{x}$$

If we let M be a n by r matrix with M_r being its rth column, the result is just:

$$\boldsymbol{h}_M(\boldsymbol{x}) = \arg\max M^T\boldsymbol{x}$$

assuming 1-based indexing and one-hot encoding, which represents $\boldsymbol{y}$ as a m by k matrix Y with each row $Y_{(i)}$ corresponding to the one-hot encoded vector of $\boldsymbol{y}_{(i)}$. We can represent the margin $\boldsymbol{\gamma}_{(r)}$ as:

$$\boldsymbol{\gamma}_{(r)} = \min_{i=1}^{m} \frac{{M_r}^T X_{(i)}}{}$$