Machine Learning Lecture Notes

Kede Ma

December 4, 2024

Math Notation

$\mathbb{R}$	the set of real numbers
$\mathbb{R}^N$	N-dimensional real space
M	the number of training samples
N	the number of input features (or data attributes) to represent a training sample
$\mathcal{D}$	the training set that consists of M training samples
x	an n-dimensional input feature vector
$x^{(i)}$	the ith input feature vector in the training set $\mathcal{D}$
$x_j^{(i)}$	the jth feature of the ith input feature vector in $\mathcal{D}$
$y^{(i)}$	the ith output label in $\mathcal{D}$ corresponding to $x^{(i)}$
$\mathcal{Y}$	the set of all output labels
$\vert\mathcal{Y}\vert$	the cardinality of $\mathcal{Y}$ (*i.e.*, the number of classes in $\mathcal{Y}$)
$\mathbb{I}[\cdot]$	the indicator function
$\mathrm{sign}(\cdot)$	the sign function
$\sum$	the summation of multiple terms
$\prod$	the product of multiple terms
$\int$	the integration with respect to continuous variables
$\frac{\partial f(z)}{\partial z_j}$	the partial derivative of $f(z)$ w.r.t. z_j, where $z = [z_1, z_2, \ldots, z_N]^T$
$\nabla f(z)$	the derivative of $f(z)$ at z, where $\nabla f(z) = [\frac{\partial f(z)}{\partial z_1}, \ldots, \frac{\partial f(z)}{\partial z_N}]^T$
$\nabla_v f(z)$	the directional derivative of f along the direction v at z
∞	infinity
$\lim\limits_{z \to a} f(z)$	the limit of $f(z)$ as z approaches a
A^T	the transpose of a matrix A
A^{-1}	the inverse of a square matrix A
A^{-T}	the inverse of the transposed A and vice versa, $A^{-T} = (A^{-1})^T = (A^T)^{-1}$
$\mathrm{tr}A$	the trace of a square matrix A
$\Vert A\Vert_F$	the Frobenius norm of a matrix A
$\perp$	perpendicular to

1 Lecture 1

1.1 Derivation of the Bayes Optimal Classifier

Proof. Given a classifier $f \in \mathcal{H}$, where $\mathcal{H}$ denotes the set of candidate classifiers that includes the best one, we first write down the 0-1 loss for a training sample (x, y):

$$\ell(f(x), y) = \mathbb{I}[f(x) \neq y] = 1 - \mathbb{I}[f(x) = y]. \tag{1}$$

The expected predicted error of f is therefore

$$\ell(f) = \mathbb{E}_{x,y}[\ell(f(x), y)], \tag{2}$$

where the expectation is taken w.r.t. the joint distribution $p(x, y)$. Expanding Eq. (2), we have

$$\begin{aligned}
\ell(f) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \ell(f(x), y) \\
&= \sum_{x \in \mathcal{X}} p(x) \left[\sum_{y \in \mathcal{Y}} p(y|x) \ell(f(x), y) \right] \\
&= \mathbb{E}_x \left[\sum_{y \in \mathcal{Y}} p(y|x) \ell(f(x), y) \right]. \qquad (3)
\end{aligned}$$

As we would like to minimize the expected loss, for each x, we have

$$\begin{aligned}
f_B &= \arg\min_{f \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x) \ell(f(x), y) \\
&= \arg\min_{f \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x)(1 - \mathbb{I}[f(x) = y]) \\
&= \arg\min_{f \in \mathcal{H}} \underbrace{\sum_{y \in \mathcal{Y}} p(y|x)}_{1} - \sum_{y \in \mathcal{Y}} p(y|x) \mathbb{I}[f(x) = y]) \\
&= \arg\min_{f \in \mathcal{H}} - \sum_{y \in \mathcal{Y}} p(y|x) \mathbb{I}[f(x) = y]) \\
&= \arg\max_{f \in \mathcal{H}} \sum_{y \in \mathcal{Y}} p(y|x) \mathbb{I}[f(x) = y] \\
&= \arg\max_{f \in \mathcal{H}} p(f(x)|x). \qquad (4)
\end{aligned}$$

Thus,

$$f_B(x) = \arg\max_{y \in \mathcal{Y}} p(y|x), \tag{5}$$

and the corresponding expected error rate is

$$\ell = 1 - \mathbb{E}_x \left[\max_{y \in \mathcal{Y}} p(y|x) \right]. \tag{6}$$

In other words, the optimal Bayes decision rule is to choose the class presenting the maximum posterior probability, given the particular observation at hand. Classifiers such as these are called Bayes Optimal Classifiers or Maximum a Posteriori classifiers. □

2 Lecture 2

2.1 Derivation of the MLE for Naive Bayes

Proof. Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1, \dots, M\}$, we write down the joint probability distribution of the data

$$\begin{aligned} p(\mathcal{D}; \theta) &= \prod_{i=1}^{M} p(x^{(i)}, y^{(i)}; \theta) \\ &= \prod_{i=1}^{M} p(y^{(i)}; \theta_1) p(x^{(i)}|y^{(i)}; \theta_2) \\ &= \prod_{i=1}^{M} p(y^{(i)}; \theta_1) \prod_{j=1}^{N} p(x_j^{(i)}|y^{(i)}; \theta_2), \end{aligned} \tag{7}$$

where the parameter vector $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$ is composed of two (non-overlapping) sub-vectors, one associated with the prior distribution $p(y^{(i)})$ and the other associated with the class conditionals $p(x^{(i)}|y^{(i)})$. When we wish to explicitly view this as a function of the parameter vector θ, we instead call it the likelihood function of the data $L(\theta)$. The principal of maximum likelihood says that we should choose θ so as to make the data as high probability as possible. That is, we should choose θ to maximize $L(\theta)$. Instead of maximizing $L(\theta)$, we can also maximize any strictly increasing function of $L(\theta)$. In particular, the derivations will be a bit simpler if we instead maximize the log likelihood

$$\ell(\theta) = \sum_{i=1}^{M} \log p(y^{(i)}; \theta_1) + \sum_{i=1}^{M} \sum_{j=1}^{N} \log p(x_j^{(i)}|y^{(i)}; \theta_2). \tag{8}$$

Assume that $y^{(i)}$ represents one of C possible classes, *i.e.*, $y^{(i)} \in \{1, 2, \dots, C\}$, and therefore can be modeled using a categorical distribution

$$p(y^{(i)}; \varphi) = \prod_{c=1}^{C} \varphi_c^{\mathbb{I}[y^{(i)}=c]}, \tag{9}$$

where $\varphi = [\varphi_1, \varphi_2, \dots, \varphi_C]^T$ and $\sum_{c=1}^{C} \varphi_c = 1$. Comparing Eqs. (8) and (9), we have $\theta_1 = \varphi$. That is, we replace θ_1 from an unspecified distribution with φ from the categorical distribution. Now we can see that if we want to estimate φ, the second part of the log-likelihood function is irrelevant, as it does not depend on φ. Note also that we have a constraint on φ, which is $\sum_{c=1}^{C} \varphi_c = 1$. We can use the method of Lagrange multipliers, which leads us to the following optimization problem

$$J(\varphi) = \sum_{i=1}^{M} \sum_{c=1}^{C} \mathbb{I}[y^{(i)} = c] \log \varphi_c + \lambda \left(\sum_{c=1}^{C} \varphi_c - 1 \right). \tag{10}$$

Taking the derivative w.r.t. φ_c and setting it to zero, we have

$$\begin{aligned} \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}{\varphi_c} + \lambda = 0, \\ \varphi_c = \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}{-\lambda}. \end{aligned} \tag{11}$$

By exploiting the fact $\sum_{c=1}^{C} \varphi_c = 1$, we have

$$-\lambda = \sum_{i=1}^{M} \sum_{c=1}^{C} \mathbb{I}[y^{(i)} = c] = M. \tag{12}$$

Therefore,

$$\varphi_c = \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}{M}. \tag{13}$$

For real valued x_j, we model it with a Gaussian distribution

$$p(x_j | y = c) = \frac{1}{\sqrt{2\pi\sigma_{j|c}^2}} \exp\left(-\frac{(x_j - \mu_{j|c})^2}{2\sigma_{j|c}^2} \right). \tag{14}$$

Here, again we replace θ_2 in Eq. (8) from an unspecified distribution with $\{\mu_{j|c}, \sigma_{j|c}^2\}$ from Gaussian distributions. If we pick out all terms in Eq. (8) that depend only on $\mu_{j|c}$, $\sigma_{j|c}^2$, we have

$$J(\mu_{j|c}, \sigma_{j|c}^2) = \sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \left(-\frac{1}{2} \log 2\pi\sigma_{j|c}^2 - \frac{(x_j^{(i)} - \mu_{j|c})^2}{2\sigma_{j|c}^2} \right). \tag{15}$$

Taking the derivative w.r.t. $\mu_{j|c}$ and setting it to zero, we have

$$\begin{aligned} \sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c](x_j^{(i)} - \mu_{j|c}) = 0 \\ \mu_{j|c} = \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] x_j^{(i)}}{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}. \end{aligned} \tag{16}$$

Taking the derivative w.r.t. $\sigma^2_{j|c}$ and setting it to zero, we have

$$\begin{aligned}
\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \left(-\frac{1}{2\sigma^2_{j|c}} + \frac{(x_j^{(i)} - \mu_{j|c})^2}{2\sigma^4_{j|c}} \right) &= 0 \\
\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \left(-1 + \frac{(x_j^{(i)} - \mu_{j|c})^2}{\sigma^2_{j|c}} \right) &= 0 \\
\sigma^2_{j|c} = \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c](x_j^{(i)} - \mu_{j|c})^2}{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}. \qquad (17)
\end{aligned}$$

□

2.2 Derivation of the MLE for Linear Discriminant Analysis

Proof. We start with the log likelihood of the data

$$\begin{aligned}
\ell(\varphi, \mu, \Sigma) &= \log \prod_{i=1}^{M} p(x^{(i)}, y^{(i)}; \varphi, \mu, \Sigma) \\
&= \log \prod_{i=1}^{M} p(x^{(i)} | y^{(i)}; \mu, \Sigma) p(y^{(i)}; \varphi) \\
&= \log \prod_{i=1}^{M} \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right) \prod_{c=1}^{C} \varphi_c^{\mathbb{I}[y^{(i)}=c]} \\
&= \sum_{i=1}^{M} \Big[-\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \\
&+ \sum_{c=1}^{C} \mathbb{I}[y^{(i)} = c] \log \varphi_c \Big]. \qquad (18)
\end{aligned}$$

Now we need to take partial derivatives w.r.t. each parameter and equate it to zero. For μ_c,

$$\frac{\partial \ell(\varphi, \mu_c, \Sigma)}{\partial \mu_c} = \sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \Sigma^{-1} (x^{(i)} - \mu_c) = 0. \qquad (19)$$

Since the nullspace of Σ^{-1} is $\{0\}$, we have

$$\begin{aligned}
\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c](x^{(i)} - \mu_c) &= 0 \\
\mu_c &= \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] x^{(i)}}{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}. \qquad (20)
\end{aligned}$$

For Σ,

$$\frac{\partial \ell(\varphi, \mu, \Sigma)}{\partial \Sigma} = \sum_{i=1}^{M} \left(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} \right) = 0$$

$$\sum_{i=1}^{M} \Sigma^{-1} \left(-I + (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} \right) = 0$$

$$\Sigma = \frac{1}{M} \sum_{i=1}^{M} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T, \tag{21}$$

where the first equation follows from the facts that

$$\frac{\partial a^T X^{-1} b}{\partial X} = -X^{-T} a b^T X^{-T} \tag{22}$$

and

$$\frac{\partial \log |X|}{\partial X} = (X^{-1})^T = X^{-1}, \tag{23}$$

for a symmetric matrix X. □

2.3 Probabilistic View of the Sigmoid Function

Let's write the general sigmoid function

$$F(x; \mu, s) = \frac{1}{1 + e^{-\frac{x-\mu}{s}}}. \tag{24}$$

For a function $F(x)$ to be a legitimate cumulative distribution function (CDF), it must satisfy four properties:

- Non-decreasing;
- Right-continuous (*i.e.*, $\lim_{x \to c^+} F(x) = F(c)$);
- $\lim_{x \to -\infty} F(x) = 0$;
- $\lim_{x \to +\infty} F(x) = 1$.

Obviously, the sigmoid function satisfies all the four properties, and the corresponding probability density function (PDF) is

$$f(x) = F'(x) = \frac{e^{-r}}{s\left(1 + e^{-r}\right)^2} \quad \text{and} \quad r = \frac{x - \mu}{s}, \tag{25}$$

which is known as the logistic distribution.

3 Lecture 3

3.1 Local and Global Optima of Convex Functions

Any locally optimal point of a convex function $f : \mathbb{R}^N \mapsto \mathbb{R}$ is (globally) optimal.

Proof. Suppose x is locally optimal and y is globally optimal with $f(y) < f(x)$. The locally optimal x means that there is a radius $R > 0$ such that

$$z \in \mathrm{dom}(f), \quad \|z - x\|_2 \leq R \Longrightarrow f(x) \leq f(z). \tag{26}$$

Now consider $z = \theta y + (1 - \theta)x$ with $\theta = R/(2\|y - x\|_2)$. First, we note that $\|y - x\|_2 > R$, otherwise it violates the assumptions that x is locally optimal and y is globally optimal. Therefore, we have $0 \leq \theta \leq 1/2$, which indicates that z is a convex combination of two feasible points and is in the $\mathrm{dom}(f)$. Note also that

$$\|z - x\|_2 = \|\theta y + (1 - \theta)x - x\|_2 = \|\theta(y - x)\|_2 = \theta\|(y - x)\|_2 = \frac{R}{2}, \tag{27}$$

which implies $f(x) \leq f(z)$ according to (26). As f is convex, we have

$$f(z) = f(\theta y + (1 - \theta)x) \leq \theta f(y) + (1 - \theta)f(x) < \theta f(x) + (1 - \theta)f(x) = f(x), \tag{28}$$

which contradicts our assumption that x is locally optimal. As a result, any locally optimal point of a convex function is globally optimal. □

3.2 Gradient Descent of Logistic Regression

In logistic function, we use the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \; z \in \mathbb{R} \tag{29}$$

to map the linear function

$$f(x) = w^T x + b \tag{30}$$

to a probability value between 0 and 1, i.e,

$$p(y = +1|x) = \sigma(f(x)) \tag{31}$$

and

$$p(y = -1|x) = 1 - \sigma(f(x)) = 1 - \frac{1}{1 + e^{-f(x)}} = \frac{e^{-f(x)}}{1 + e^{-f(x)}} = \frac{1}{1 + e^{f(x)}} = \sigma(-f(x)), \tag{32}$$

where we denote a negative example by -1. Combining the above two equations, we have

$$p(y|x) = \sigma(yf(x)). \tag{33}$$

An interesting property of the sigmoid function is that its gradient can be written as a function of itself:

$$\begin{aligned}\sigma'(z) &= \frac{-(1+e^{-z})'}{(1+e^{-z})^2} \\ &= \frac{e^{-z}}{(1+e^{-z})^2} \\ &= \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} \\ &= \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right) \\ &= \sigma(z) \cdot (1-\sigma(z)). \end{aligned} \tag{34}$$

Let's start by working with just one training example (x, y), and learn the function parameters $\{w, b\}$ by minimizing the negative log likelihood:

$$\ell(w,b) = -\log p(y|x; w, b) = -\log\left(\sigma(yf(x))\right) = \log\left(1 + \exp\left(-y(w^T x + b)\right)\right). \tag{35}$$

Taking the partial derivative of ℓ w.r.t. the jth element of the parameter vector, w_j, we have:

$$\begin{aligned}\frac{\partial}{\partial w_j}\ell(w,b) &= -\frac{1}{\sigma(yf(x))} \cdot \frac{\partial}{\partial w_j}\sigma(yf(x)) \\ &= -\frac{1}{\sigma(yf(x))}\sigma(yf(x))\left(1-\sigma(yf(x))\right) \cdot \frac{\partial}{\partial w_j} yf(x) \\ &= \left(\sigma(yf(x)) - 1\right) \cdot \frac{\partial}{\partial w_j} yf(x) \\ &= \left(\sigma(yf(x)) - 1\right) \cdot \frac{\partial}{\partial w_j} y(w^T x + b) \\ &= \left(\sigma(yf(x)) - 1\right) y x_j. \end{aligned} \tag{36}$$

This therefore gives us the stochastic gradient descent rule:

$$w_j^{(t+1)} = w_j^{(t)} - \alpha\left(\sigma\left(y\left((w^{(t)})^T x + b^{(t)}\right)\right) - 1\right) y x_j, \tag{37}$$

from which we can easily derive the gradient descent rule:

$$w_j^{(t+1)} = w_j^{(t)} - \alpha \frac{1}{M}\sum_{i=1}^{M}\left(\sigma\left(y^{(i)}\left((w^{(t)})^T x^{(i)} + b^{(t)}\right)\right) - 1\right) y^{(i)} x_j^{(i)}. \tag{38}$$

Finally, the corresponding vectorized gradient descent rule is

$$w^{(t+1)} = w^{(t)} - \alpha \frac{1}{M}\sum_{i=1}^{M}\left(\sigma\left(y^{(i)}\left((w^{(t)})^T x^{(i)} + b^{(t)}\right)\right) - 1\right) y^{(i)} x^{(i)}. \tag{39}$$

The derivative of ℓ w.r.t. the bias term b can be derived similarly according to Eq. (36).

It is important to note that the above derivation is based on the class set $\mathcal{Y} = \{+1, -1\}$, where we denote a negative example by -1. It is perfectly fine to work with the class set $\mathcal{Y} = \{+1, 0\}$, and denote a negative example by 0. In this case, we have

$$p(y = +1|x) = \sigma(f(x)) \tag{40}$$

and

$$p(y = 0|x) = 1 - \sigma(f(x)). \tag{41}$$

Fortunately, we are still able to combine the above two equations:

$$p(y|x) = \sigma(f(x))^y \cdot (1 - \sigma(f(x)))^{1-y}. \tag{42}$$

We again start by working with just one training example (x, y), and learn the function parameters $\{w, b\}$ by minimizing the negative log likelihood:

$$\begin{aligned}\ell(w, b) &= -\log p(y|x; w, b) = -\log\left(\sigma(f(x))^y \cdot (1 - \sigma(f(x)))^{1-y}\right)\\ &= -y \log \sigma(f(x)) - (1 - y) \log(1 - \sigma(f(x)))\\ &= -y \log \sigma(w^T x + b) - (1 - y) \log(1 - \sigma(w^T x + b)).\end{aligned} \tag{43}$$

Take derivative of ℓ w.r.t. w_j:

$$\begin{aligned}\frac{\partial}{\partial w_j}\ell(w, b) &= \left(-y\frac{1}{\sigma(f(x))} + (1 - y)\frac{1}{1 - \sigma(f(x))}\right) \cdot \frac{\partial}{\partial w_j}\sigma(f(x))\\ &= \left(-y\frac{1}{\sigma(f(x))} + (1 - y)\frac{1}{1 - \sigma(f(x))}\right) \sigma(f(x))(1 - \sigma(f(x))) \cdot \frac{\partial}{\partial w_j} f(x)\\ &= (-y(1 - \sigma(f(x))) + (1 - y)\sigma(f(x))) \cdot \frac{\partial}{\partial w_j} f(x)\\ &= (\sigma(f(x)) - y) \cdot \frac{\partial}{\partial w_j}(w^T x + b)\\ &= (\sigma(f(x)) - y)\, x_j,\end{aligned} \tag{44}$$

which is equivalent to Eq. (36) for both positive and negative examples.

Last, we may add an ℓ_2-term to penalize large values in the parameter vector w. The loss function on a single training example (x, y) defined in Eq. (35) now becomes

$$\ell(w, b) = \log\left(1 + \exp\left(-y(w^T x + b)\right)\right) + \frac{1}{C} w^T w, \tag{45}$$

where C is a hyperparameter. The partial derivative of ℓ w.r.t. w_j defined in Eq. (36) changes accordingly:

$$\frac{\partial}{\partial w_j}\ell(w, b) = (\sigma(y f(x)) - 1)\, y x_j + \frac{2}{C} w_j. \tag{46}$$

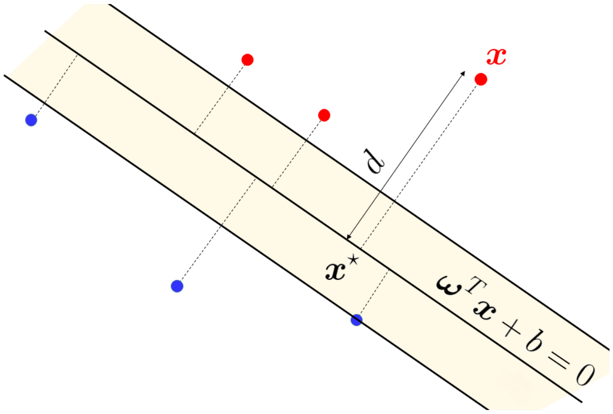

Figure 1: Computation of the margin d.

3.3 Computation of the Margin in SVM

Proof. First, note that $w/\|w\|_2$ is a unit-length vector pointing in the same direction as w. We are able to compute the projection of $x^{(i)}$ on the hyperplane $w^Tx+b=0$ as $x^{(i)}-d^{(i)}\cdot w/\|w\|_2$, where we assume $y^{(i)}=1$. Since all points on the hyperplane satisfy the equation $w^Tx+b=0$, we have

$$w^T\left(x^{(i)}-d^{(i)}\frac{w}{\|w\|_2}\right)+b=0. \tag{47}$$

Solving for $d^{(i)}$ yields

$$d^{(i)}=\frac{w^Tx^{(i)}+b}{\|w\|_2}, \tag{48}$$

where we use the definitions of $\|w\|_2=\sqrt{\sum_j w_j^2}$ and $w^Tw=\|w\|_2^2=\sum_j w_j^2$. □

If $y^{(i)}=-1$, the projection of $x^{(i)}$ on the hyperplane becomes $x^{(i)}+d^{(i)}\cdot w/\|w\|_2$, and we have the following equation

$$w^T\left(x^{(i)}+d^{(i)}\frac{w}{\|w\|_2}\right)+b=0. \tag{49}$$

Solving for $d^{(i)}$ yields

$$d^{(i)}=\frac{-(w^Tx^{(i)}+b)}{\|w\|_2}. \tag{50}$$

Unifying Eq. (48) and Eq. (50) using the label information $y^{(i)}$, we have

$$d^{(i)}=\frac{y^{(i)}(w^Tx^{(i)}+b)}{\|w\|_2}, \tag{51}$$

as desired.

3.4 Matrix Derivatives[1]

For a function $f : \mathbb{R}^{M \times N} \mapsto \mathbb{R}$ mapping from an M-by-N matrix to a real number, we define the derivative of f with respect to its input A to be:

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \cdots & \frac{\partial f(A)}{\partial A_{1N}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{M1}} & \cdots & \frac{\partial f(A)}{\partial A_{MN}} \end{bmatrix}. \tag{52}$$

Thus, the gradient $\nabla_A f(A)$ is itself an M-by-N matrix, whose (i, j)-element is $\frac{\partial f}{\partial A_{ij}}$. For example, suppose $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ is a 2-by-2 matrix, and the function $f : \mathbb{R}^{2\times 2} \mapsto \mathbb{R}$ is given by

$$f(A) = A_{11}^2 A_{12} + 2A_{21} + A_{22}^3. \tag{53}$$

We then have

$$\nabla_A f(A) = \begin{bmatrix} 2A_{11}A_{12} & A_{11}^2 \\ 2 & 3A_{22}^2 \end{bmatrix}. \tag{54}$$

4 Lecture 4

4.1 Dual Form of SVM with Soft Margin

Proof. We first write the primal form of SVM with soft margin

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{M} \xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \dots, M, \\ & \xi_i \geq 0, \quad i = 1, \dots, M, \end{aligned} \tag{55}$$

which may be rewritten in standard form as

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{M} \xi_i \\ \text{s.t.} \quad & -[y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i] \leq 0, \quad i = 1, \dots, M, \\ & -\xi_i \leq 0, \quad i = 1, \dots, M. \end{aligned} \tag{56}$$

We then form the Lagrangian of the above problem

$$\ell(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^T w + C \sum_{i=1}^{M} \xi_i - \sum_{i=1}^{M} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^{M} \beta_i \xi_i \tag{57}$$

$$= \frac{1}{2} w^T w - \sum_{i=1}^{M} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{M} \alpha_i y^{(i)} b + \sum_{i=1}^{M} \xi_i (C - \alpha_i - \beta_i) + \sum_{i=1}^{M} \alpha_i, \tag{58}$$

[1] Inspired by Andrew Ng's lecture notes.

where α_i's and β_i's are the Lagrange multipliers (constrained to be ≥ 0). We obtain the Lagrange dual function $g(\alpha, \beta)$ by minimizing the Lagrangian $\ell(w, b, \xi, \alpha, \beta)$ w.r.t. the primal variables w, b, and ξ. To do that, we take the partial derivatives of ℓ w.r.t. w, b, ξ and set them to zero

$$\nabla_w \ell = \nabla_w \left(\frac{1}{2} w^T w - \sum_{i=1}^{M} \alpha_i y^{(i)} w^T x^{(i)} \right) = w - \sum_{i=1}^{M} \alpha_i y^{(i)} x^{(i)} = 0, \tag{59}$$

$$\nabla_b \ell = \nabla_b \left(- \sum_{i=1}^{M} \alpha_i y^{(i)} b \right) = - \sum_{i=1}^{M} \alpha_i y^{(i)} = 0, \tag{60}$$

$$\nabla_{\xi_i} \ell = \nabla_{\xi_i} \left(\sum_{i=1}^{M} \xi_i (C - \alpha_i - \beta_i) \right) = C - \alpha_i - \beta_i = 0, \quad i = 1, \ldots, M. \tag{61}$$

Plugging Eq. (59), Eq. (60), and Eq. (61) back into the Lagrangian (Eq. (58)), we obtain

$$g(\alpha) = \frac{1}{2} \left(\sum_{i=1}^{M} \alpha_i y^{(i)} x^{(i)} \right)^T \left(\sum_{j=1}^{M} \alpha_j y^{(j)} x^{(j)} \right) - \sum_{i=1}^{M} \alpha_i y^{(i)} \left(\sum_{j=1}^{M} \alpha_j y^{(j)} x^{(j)} \right)^T x^{(i)} + \sum_{i=1}^{M} \alpha_i \tag{62}$$

$$= \sum_{i=1}^{M} \alpha_i + \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} - \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \tag{63}$$

$$= \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}. \tag{64}$$

In Eq. (63), we make use of the distributive property of matrix product: $(a+b)^T(c+d) = (a^T + b^T)(c+d) = a^T c + a^T d + b^T c + b^T d$, for $a, b, c, d \in \mathbb{R}^N$. Also note that $(x^{(i)})^T x^{(j)} = (x^{(j)})^T x^{(i)}$. Putting $g(\alpha)$ together with the constraints, we obtain the following dual optimization problem:

$$\begin{aligned}
\max_{\alpha, \beta} \quad & \sum_{i=1}^{M} \alpha_i - \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)} \\
\text{s.t.} \quad & \sum_{i=1}^{M} \alpha_i y^{(i)} = 0, \\
& C - \alpha_i - \beta_i = 0, \quad i = 1, \ldots, M, \\
& \alpha_i \geq 0, \quad i = 1, \ldots, M, \\
& \beta_i \geq 0, \quad i = 1, \ldots, M.
\end{aligned}$$

It is straightforward to show that

$$\begin{cases} C - \alpha_i - \beta_i = 0 \\ \alpha_i \geq 0 \\ \beta_i \geq 0 \end{cases} \iff \begin{cases} \alpha_i \geq 0 \\ \alpha_i \leq C. \end{cases} \tag{65}$$

The dual optimization problem can then be simplified as

$$\begin{aligned}\max_{\alpha} \quad & \sum_{i=1}^{M} \alpha_i - \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^T x^{(j)} \\ \text{s.t.} \quad & \sum_{i=1}^{M} \alpha_i y^{(i)} = 0, \\ & 0 \le \alpha_i \le C, \quad i = 1, \ldots, M.\end{aligned}$$

Once α_i is determined, we are able to compute $\beta_i = C - \alpha_i$ accordingly. □

4.2 Complementary Slackness

Assuming that strong duality holds, $w^\star$ is primal optimal, and $(\lambda^\star, \nu^\star)$ are dual optimal, we have

$$\begin{aligned} f_0(w^\star) = p^\star = d^\star = g(\lambda^\star, \nu^\star) &= \inf_w \left(f_0(w) + \sum_{i=1}^{r} \lambda_i^\star f_i(w) + \sum_{i=1}^{s} \nu_i^\star h_i(w) \right) && (66) \\ &\le f_0(w^\star) + \sum_{i=1}^{r} \lambda_i^\star f_i(w^\star) + \sum_{i=1}^{s} \nu_i^\star h_i(w^\star) && (67) \\ &\le f_0(w^\star). && (68)\end{aligned}$$

Hence, the two inequalities hold with equality, which implies

- $w^\star$ not only minimizes $f_0(w)$, but also minimizes the Lagrangian $L(w, \lambda^\star, \nu^\star)$ with dual optimal Lagrange multiplier $\lambda^\star$ and $\nu^\star$.
- $\lambda_i^\star f_i(w^\star) = 0$ for $i = 1, \ldots, M$ (known as **complementary slackness**):

$$\lambda_i^\star > 0 \Longrightarrow f_i(w^\star) = 0, \qquad f_i(w^\star) < 0 \Longrightarrow \lambda_i^\star = 0. \tag{69}$$

In the context of SVM with soft margin, we have the following complementary slackness

$$\begin{aligned} \alpha_i^\star \left[y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) - 1 + \xi_i^\star \right] = 0,\ i = 1, \ldots, M, && (70) \\ \beta_i^\star \xi_i^\star = 0,\ i = 1, \ldots, M. && (71)\end{aligned}$$

Combining with **primal feasibility**

- $y^{(i)}((w^\star)^T x^{(i)} + b^\star) \ge 1 - \xi_i^\star$,
- $\xi_i^\star \ge 0$,

and **dual feasibility**

- $C - \alpha_i^\star - \beta_i^\star = 0$,
- $\alpha_i^\star \ge 0$,

- $\beta_i^\star \geq 0$,

we have

- if $\alpha^\star = 0$

$$\begin{cases} \alpha_i^\star = 0 \\ C - \alpha_i^\star - \beta_i^\star = 0 \end{cases} \implies \beta_i^\star = C, \tag{72}$$

$$\begin{cases} \beta_i^\star = C \\ \beta_i^\star \xi_i^\star = 0 \end{cases} \implies \xi_i^\star = 0, \tag{73}$$

$$\begin{cases} \xi_i^\star = 0 \\ y^{(i)}((w^\star)^T x^{(i)} + b^\star) \geq 1 - \xi_i^\star \end{cases} \implies y^{(i)}((w^\star)^T x^{(i)} + b^\star) \geq 1; \tag{74}$$

- if $\alpha^\star = C$

$$\begin{cases} \alpha_i^\star = C \\ \alpha_i^\star \left[y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) - 1 + \xi_i^\star \right] = 0 \end{cases} \implies y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) - 1 + \xi_i^\star = 0, \tag{75}$$

$$\begin{cases} y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) - 1 + \xi_i^\star = 0 \\ \xi_i^\star \geq 0 \end{cases} \implies y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) \leq 1; \tag{76}$$

- if $0 < \alpha^\star < C$ (implying $0 < \beta^\star < C$)

$$\begin{cases} 0 < \alpha^\star < C \\ C - \alpha_i^\star - \beta_i^\star = 0 \end{cases} \implies \beta_i^\star > 0, \tag{77}$$

$$\begin{cases} 0 < \alpha^\star < C \\ \alpha_i^\star \left[y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) - 1 + \xi_i^\star \right] = 0 \end{cases} \implies y^{(i)} \left((x^{(i)})^T w^\star + b^\star \right) - 1 + \xi_i^\star = 0, \tag{78}$$

$$\begin{cases} \beta_i^\star > 0 \\ \beta_i^\star \xi_i^\star = 0 \end{cases} \implies \xi_i^\star = 0, \tag{79}$$

$$\begin{cases} \xi_i^\star = 0 \\ y^{(i)}((w^\star)^T x^{(i)} + b^\star) - 1 + \xi_i^\star = 0 \end{cases} \implies y^{(i)}((w^\star)^T x^{(i)} + b^\star) = 1. \tag{80}$$

4.3 Optimal Bias Term in Dual-Form SVM with Soft-Margin

After the dual optimization problem is solved, the optimal bias term $b^\star$ can be calculated by using training samples $(x, y) \in \mathcal{U} \subset \mathcal{D}$, whose Lagrange multipliers α's are *unbounded* (*i.e.*, $0 < \alpha < C$), satisfying

$$\begin{aligned} y((w^\star)^T x + b^\star) &= 1 \\ y^2((w^\star)^T x + b^\star) &= y \\ (w^\star)^T x + b^\star &= y \\ b^\star &= y - (w^\star)^T x \end{aligned} \tag{81}$$

In Eq. (81), since $y \in \{-1, 1\}$, $y^2 = 1$. Averaging all training samples in $\mathcal{U}$, we have

$$b^\star = \frac{1}{|\mathcal{U}|} \sum_{(x,y)\in\mathcal{U}} (y - (w^\star)^T x). \tag{82}$$

4.4 Feature Mappings ϕ for Polynomial Kernels

The inhomogeneous polynomial kernel is defined as

$$K(x, z) = \phi(x)^T \phi(z) = (x^T z + c)^d, \tag{83}$$

where $x, z \in \mathbb{R}^N$ are features in the input space, d is the degree of the polynomial, and $c \geq 0$ is a constant. If $c = 0$, $K(x, z)$ reduces to the homogeneous polynomial kernel.

Let $k_0, k_1, \ldots, k_N$ denote non-negative integers, such that $\sum_{j=0}^N k_j = d$. The multinomial expansion[2] of the inhomogeneous kernel is then given as

$$\begin{aligned} K(x, z) = (x^T z + c)^d &= (c + \sum_{j=1}^N x_j z_j)^d = (c + x_1 z_1 + \ldots + x_N z_N)^d \\ &= \sum_{k_0+\ldots+k_N=d} \binom{d}{k_0, \ldots, k_N} c^{k_0} (x_1 z_1)^{k_1} \ldots (x_N z_N)^{k_N} \\ &= \sum_{k_0+\ldots+k_N=d} \binom{d}{k_0, \ldots, k_N} c^{k_0} (x_1^{k_1} \ldots x_N^{k_N})(z_1^{k_1} \ldots z_N^{k_N}) \\ &= \sum_{k_0+\ldots+k_N=d} \left(\sqrt{\binom{d}{k_0, \ldots, k_N} c^{k_0}} \prod_{j=1}^N x_j^{k_j} \right) \left(\sqrt{\binom{d}{k_0, \ldots, k_N} c^{k_0}} \prod_{j=1}^N z_j^{k_j} \right). \end{aligned} \tag{84}$$

Then, the mapping $\phi : \mathbb{R}^N \mapsto \mathbb{R}^L$ (L is the dimension of the new feature space) is given as the vector

$$\phi(x) = \left[\ldots, \sqrt{\binom{d}{k_0, \ldots, k_N} c^{k_0}} \prod_{j=1}^N x_j^{k_j}, \ldots \right]^T, \tag{85}$$

[2] https://en.wikipedia.org/wiki/Multinomial_theorem

where $(k_0, \ldots, k_N)$ ranges over all the possible assignments, such that $k_j \geq 0$ for $0 \leq j \leq N$ and $\sum_{j=0}^{N} k_j = d$. This is a classic problem in combinatorics: what is the number of ways to write d as an (ordered) sum of $N+1$ *non-negative* integers? This is equivalent to the problem of finding the number of ways to write $d+N+1$ as an (ordered) sum of $N+1$ *positive* integers (we essentially add each k_j by one to make it strictly positive, such that $\sum_{j=0}^{N}(k_j+1) = d+N+1$). With the aid of stars and bars[3], it is straightforward to show that the dimensionality of the expanded feature space is

$$L = \binom{d+N+1-1}{N+1-1} = \binom{d+N}{N} = \binom{d+N}{d+N-N} = \binom{d+N}{d}. \tag{86}$$

4.5 Feature Mappings ϕ for Gaussian Kernels

It is interesting to note that the feature space for a Gaussian kernel has infinite dimensionality. To see this, recall that the exponential function[4] can be written as the infinite expansion

$$f(\theta) = e^{\theta} = \sum_{d=0}^{\infty} \frac{\theta^d}{d!} = 1 + \theta + \frac{1}{2!}\theta^2 + \frac{1}{3!}\theta^3 + \ldots \tag{87}$$

Further, noting that $\|x-z\|_2^2 = \|x\|_2^2 + \|z\|_2^2 - 2x^T z$, we can rewrite the Gaussian kernel as follows

$$\begin{aligned} K(x,z) &= \exp\left\{-\gamma\|x-z\|_2^2\right\} \\ &= \exp\left\{-\gamma\|x\|_2^2\right\} \cdot \exp\left\{-\gamma\|z\|_2^2\right\} \cdot \exp\left\{2\gamma x^T z\right\}. \end{aligned} \tag{88}$$

In particular, the last term is given as the infinite expansion

$$\exp\left\{2\gamma x^T z\right\} = \sum_{d=0}^{\infty} \frac{(2\gamma)^d}{d!}(x^T z)^d = 1 + (2\gamma)x^T z + \frac{(2\gamma)^2}{2!}(x^T z)^2 + \ldots \tag{89}$$

Using the multinomial expansion of $(x^T z)^d$ (refer to Section 4.4), we can write the Gaussian kernel as

$$\begin{aligned} &K(x,z) \\ &= \exp\left\{-\gamma\|x\|_2^2\right\} \exp\left\{-\gamma\|z\|_2^2\right\} \left(\sum_{d=0}^{\infty} \frac{(2\gamma)^d}{d!} \left(\sum_{k_1+\ldots+k_N=d} \binom{d}{k_1,\ldots,k_N} \prod_{j=1}^{N} (x_j z_j)^{k_j} \right) \right) \\ &= \sum_{d=0}^{\infty} \sum_{k_1+\ldots+k_N=d} \left(\sqrt{\omega} \exp\left\{-\gamma\|x\|_2^2\right\} \prod_{j=1}^{N} (x_j)^{k_j} \right) \left(\sqrt{\omega} \exp\left\{-\gamma\|z\|_2^2\right\} \prod_{j=1}^{N} (z_j)^{k_j} \right) \end{aligned} \tag{90}$$

where

$$\omega = \frac{(2\gamma)^d}{d!} \binom{d}{k_1,\ldots,k_N}. \tag{91}$$

[3] https://en.wikipedia.org/wiki/Stars_and_bars_(combinatorics)

[4] https://en.wikipedia.org/wiki/Exponential_function

Therefore, the mapping $\phi : \mathbb{R}^N \mapsto \mathbb{R}^\infty$ is given as the vector

$$\phi(x) = \left[\ldots, \sqrt{\frac{(2\gamma)^d}{d!}\binom{d}{k_1, \ldots, k_N}} \exp\left\{-\gamma\|x\|_2^2\right\} \prod_{j=1}^{N} (x_j)^{k_j}, \ldots\right]^T . \tag{92}$$

Since ϕ maps the input space into an infinite dimensional feature space, we obviously cannot compute $\phi(x)$, yet computing the Gaussian kernel $K(x, z)$ is straightforward and efficient.

5 Lecture 5

5.1 Ordinary Least Squares in Matrix Form

Recall that

$$X = \begin{bmatrix} — & (x^{(1)})^T & — \\ — & (x^{(2)})^T & — \\ & \vdots & \\ — & (x^{(M)})^T & — \end{bmatrix} \tag{93}$$

is an M-by-$N+1$ matrix containing the training examples in its rows (where $x_0^{(i)} = 1$ for $i = 1, \ldots M$) and

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{bmatrix} \tag{94}$$

is an M-by-1 vector containing all the output values from the training set. We can easily verify that

$$y - Xw = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{bmatrix} - \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(M)})^T \end{bmatrix} w = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{bmatrix} - \begin{bmatrix} (x^{(1)})^T w \\ (x^{(2)})^T w \\ \vdots \\ (x^{(M)})^T w \end{bmatrix} = \begin{bmatrix} y^{(1)} - (x^{(1)})^T w \\ y^{(2)} - (x^{(2)})^T w \\ \vdots \\ y^{(M)} - (x^{(M)})^T w \end{bmatrix}, \tag{95}$$

where w is an $N+1$-dimensional parameter vector to be estimated with $w_0 = b$ (the bias term). Using the fact that for a vector z, $z^T z = \sum_i z_i^2$, we have

$$\frac{1}{M}(y - Xw)^T(y - Xw) = \frac{1}{M}\sum_{i=1}^{M} \left(y^{(i)} - (x^{(i)})^T w\right)^2 . \tag{96}$$

That is, we have successfully written the mean squared error (MSE) of linear regression on the training set $\mathcal{D}$ in matrix form, which is denoted by $\ell(w)$. Taking the derivative of $\ell(w)$

w.r.t. w, we have

$$\begin{aligned}
\nabla_w \ell(w) &= \frac{1}{M}\nabla_w (y - Xw)^T(y - Xw) \\
&= \frac{1}{M}\nabla_w (y^Ty - y^TXw - w^TX^Ty + w^TX^TXw) && (97) \\
&= \frac{1}{M}\nabla_w (y^Ty - y^TXw - y^TXw + w^TX^TXw) && (98) \\
&= \frac{1}{M}\nabla_w (y^Ty - 2y^TXw + w^TX^TXw) \\
&= \frac{1}{M}\nabla_w (w^TX^TXw - 2y^TXw) && (99) \\
&= \frac{2}{M}(X^TXw - X^Ty). && (100)
\end{aligned}$$

In Eq. (97), we make use of one property of matrix transpose[5] $(AB)^T = B^TA^T$. In Eq. (98), we exploit the fact that $w^TX^Ty \in \mathbb{R}$ is scalar and the transpose of a scalar is just itself, *i.e.*, $a^T = a$ for all $a \in \mathbb{R}$. In Eq. (99), since y^Ty does not depend on the parameter vector w, $\nabla_w y^Ty = 0$. In Eq. (100), we rely on two identities of matrix derivatives[6]

$$\nabla_w w^TX^TXw = (X^TX + (X^TX)^T)w = 2X^TXw \tag{101}$$

and

$$\nabla_w y^TXw = (y^TX)^T = X^Ty. \tag{102}$$

Setting the derivative in Eq. (100) to zero, we obtain the **normal equations**

$$\begin{aligned}
&\frac{2}{M}(X^TXw - X^Ty) = 0 \\
&X^TXw = X^Ty. && (103)
\end{aligned}$$

Thus, the value of w that minimizes $\ell(w)$ (denoted by $w^\star$) is given in closed form by

$$w^\star = (X^TX)^{-1}X^Ty, \tag{104}$$

where the "-1" in the superscript denotes the inverse of a square matrix[7].

5.2 A Probabilistic View of Linear Regression

The output y is from a deterministic function with additive Gaussian noise:

$$y = Xw + \epsilon, \text{ where } p(\boldsymbol{\epsilon}; \sigma^2) = \mathcal{N}(\epsilon; 0, \sigma^2 I). \tag{105}$$

[5] `https://en.wikipedia.org/wiki/Transpose`

[6] Please refer to Equation (81) and Equation (69) in The Matrix Cookbook at `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

[7] `https://en.wikipedia.org/wiki/Invertible_matrix`

Equivalently,

$$p(y|X;w,\sigma^2) = p(y - Xw; \sigma^2) = \mathcal{N}(y - Xw; 0, \sigma^2 I) = \mathcal{N}(y; Xw, \sigma^2 I). \tag{106}$$

Recall the multivariate Gaussian distribution

$$\mathcal{N}(z;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^M|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right), \tag{107}$$

and we have

$$\begin{aligned} p(y|X;w,\sigma^2) &= \frac{1}{\sqrt{(2\pi)^M|\sigma^2 I|}} \exp\left(-\frac{1}{2}(y - Xw)^T(\sigma^2 I)^{-1}(y - Xw)\right) \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^M}} \exp\left(-\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)\right). \end{aligned} \tag{108}$$

where M is the dimension of y, *i.e.*, the number of training samples. In Eq. (108), we exploit the facts that $(\sigma^2 I)^{-1} = \frac{1}{\sigma^2} I$ (the inverse of a diagonal matrix is obtained by replacing each element in the diagonal with its reciprocal), $a^T I b = a^T b$ for any two vectors a and b of the same dimension, and $|\sigma^2 I| = (\sigma^2)^M |I| = (\sigma^2)^M$ (properties of matrix determinant[8]). Maximizing the above log conditional likelihood, we have

$$\begin{aligned} &\max_{w,\sigma^2} \log p(y|X;w,\sigma^2) \\ =&\max_{w,\sigma^2} -\frac{M}{2}\log 2\pi - \frac{M}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw) \\ =&\min_{w,\sigma^2} \frac{M}{2}\log\sigma^2 + \frac{1}{2\sigma^2}(y - Xw)^T(y - Xw). \end{aligned} \tag{109}$$

In Eq. (109), we remove the constants $(-\frac{M}{2}\log 2\pi)$ that are irrelevant in the estimation of w and σ^2, and change the maximization problem into a minimization one by flipping the sign of the objective function. Taking the partial derivative of $\log p(y|X;w,\sigma^2)$ w.r.t. w and setting it to zero, we have

$$w^\star = (X^T X)^{-1} X^T y. \tag{110}$$

Taking the partial derivative of $\log p(y|X;w^\star,\sigma^2)$ (with the optimal $w^\star$ plugged in) w.r.t. σ^2 and setting it to zero, we have

$$\begin{aligned} &\frac{M}{2(\sigma^\star)^2} - \frac{1}{2(\sigma^\star)^4}(y - Xw^\star)^T(y - Xw^\star) = 0 \\ &(\sigma^\star)^2 = \frac{1}{M}(y - Xw^\star)^T(y - Xw^\star), \end{aligned} \tag{111}$$

which is the MSE of the target outputs around the regression function (and can be interpreted as a form of uncertainty).

[8] `https://en.wikipedia.org/wiki/Determinant`

5.3 On Invertibility of Normal Equations

We first note that $X^T X$ is an $(N+1) \times (N+1)$ matrix and X is an $M \times (N+1)$ matrix. We will rely on the concepts of **matrix rank**[9] and **nullspace**[10] to understand the invertibility of $X^T X$. In linear algebra, the rank of a matrix $A \in \mathbb{R}^{M \times N}$ is the dimension of the linear space generated (or spanned) by its columns. This corresponds to the maximal number of **linearly independent** columns of A. The nullspace of A is a linear subspace containing all vectors $\{v | Av = 0\}$. We have a beautiful rank–nullity theorem[11] to connect these two concepts:

$$\text{rank} A + \dim(\text{nullspace}(A)) = N, \tag{112}$$

where $\dim(\cdot)$ computes the dimension of the nullspace of A. If $A \in \mathbb{R}^{N \times N}$ is a square invertible matrix, all its columns must be linearly independent, *i.e.*, $\text{rank} A = N$ and vice versa. Back to our discussion, if $X^T X$ is not inverible, the rank of $X^T X$ is less than $N+1$, *i.e.*, $\text{rank} X^T X < N+1$.

We first show $\text{rank} X^T X = \text{rank} X$ by proving that they have exactly the same nullspace (see Eq. (112)). For one direction, if $Xv = 0$ for some v, then clearly $X^T X v = 0$. For the other direction, if $X^T X v = 0$, then $v^T X^T X v = (Xv)^T (Xv) \overset{u=Xv}{=} \sum_{j=1}^{M} u_j^2 = 0$, and it follows that $u = Xv = 0$. This implies X and $X^T X$ have the same nullspace, hence the same rank.

If $M < N+1$, that is, we simply have too many features with respect to the number of training examples, $\text{rank} X^T X = \text{rank} X \leq \min\{M, N+1\} = M < N+1$, leading to non-invertibility[12]. In this case, we may increase the number of training examples.

If $M \geq N+1$ and $\text{rank} X < N+1$, *i.e.*, the feature set is redundant (linearly dependent), we have $\text{rank} X^T X = \text{rank} X < N+1$, leading to non-invertibility. In this case, try to remove redundant features.

5.4 Closed-Form Solution for Lasso Regression under the Orthogonal Design

Proof. Under the orthogonal design (*i.e.*, $X^T X = I$), the closed form expression for standard linear regression is

$$w^{\text{LS}} = (X^T X)^{-1} X^T y = X^T y. \tag{113}$$

[9] `https://en.wikipedia.org/wiki/Rank_(linear_algebra)`

[10] `https://en.wikipedia.org/wiki/Kernel_(linear_algebra)`

[11] `https://en.wikipedia.org/wiki/Rank-nullity_theorem`

[12] The rank of an $M \times N$ matrix is a nonnegative integer and cannot be greater than either M or N. That is, $\text{rank}(A) \leq \min\{M, N\}$.

We relate the lasso objective to w^{LS} by

$$\begin{aligned}
w^{\mathrm{LR}} &= \arg\min_w \frac{1}{2}\|Xw-y\|_2^2 + \lambda\|w\|_1 \\
&= \arg\min_w \frac{1}{2}(Xw-y)^T(Xw-y) + \lambda\|w\|_1 \\
&= \arg\min_w \frac{1}{2}(w^TX^TXw - 2w^TX^Ty + y^Ty) + \lambda\|w\|_1 \\
&= \arg\min_w \frac{1}{2}(w^Tw - 2w^Tw^{\mathrm{LS}} + y^Ty) + \lambda\|w\|_1 && (114) \\
&= \arg\min_w \frac{1}{2}(w^Tw - 2w^Tw^{\mathrm{LS}}) + \lambda\|w\|_1 && (115) \\
&= \arg\min_w \sum_{j=1}^{n}(\frac{1}{2}w_j^2 - w_j^{\mathrm{LS}}w_j + \lambda|w_j|). && (116)
\end{aligned}$$

In Eq. (114), we make use of the assumption $X^TX = I$ and the notation $w^{\mathrm{LS}} = X^Ty$. In Eq. (115), we ignore the term y^Ty as it does not depend on the parameter vector w we wish to estimate. In Eq. (116), we express our objective function as a sum of objectives, each corresponding to a separate parameter w_j, and can be solved individually. Fixing a certain j, we want to minimize

$$\ell_j(w_j) = \frac{1}{2}w_j^2 - w_j^{\mathrm{LS}}w_j + \lambda|w_j|. \tag{117}$$

We note that if $w_j^{\mathrm{LS}} > 0$, we must have $w_j \geq 0$ (*i.e.*, the second term $w_j^{\mathrm{LS}}w_j$ should be nonnegative). Otherwise we could simply flip the sign of w_j and get a lower value for the objective function. Likewise, if $w_j^{\mathrm{LS}} < 0$, we must choose $w_j \leq 0$.
Case I: $w_j^{\mathrm{LS}} > 0$. Since $w_j \geq 0$,

$$\ell_j = \frac{1}{2}w_j^2 - w_j^{\mathrm{LS}}w_j + \lambda w_j. \tag{118}$$

Differentiating ℓ_j w.r.t. w_j and setting it to zero, we have

$$\begin{aligned}
w_j^{\mathrm{LR}} &= \max(0, w_j^{\mathrm{LS}} - \lambda) && (119) \\
&= \mathrm{sign}(w_j^{\mathrm{LS}})\max(0, |w_j^{\mathrm{LS}}| - \lambda). && (120)
\end{aligned}$$

In Eq. (119), we constrain w_j^{LR} to be great than or equal to zero (remember the assumptions in Case I we make) by adding a max operation. In Eq. (120), since $w_j^{\mathrm{LS}} > 0$, $\mathrm{sign}(w_j^{\mathrm{LS}}) = 1$ (refer to the definition of the sign function on Wikipedia [13]) and $|w_j^{\mathrm{LS}}| = w_j^{\mathrm{LS}}$.
Case II: $w_j^{\mathrm{LS}} < 0$. Since $w_j \leq 0$,

$$\ell_j = \frac{1}{2}w_j^2 - w_j^{\mathrm{LS}}w_j - \lambda w_j. \tag{121}$$

[13] https://en.wikipedia.org/wiki/Sign_function

Differentiating ℓ_j w.r.t. w_j and setting it to zero, we get

$$\begin{aligned} w_j^{\mathrm{LR}} &= \min(0, w_j^{\mathrm{LS}} + \lambda) && (122)\\ &= \min(0, -(-w_j^{\mathrm{LS}} - \lambda))\\ &= \min(0, -(|w_j^{\mathrm{LS}}| - \lambda)) && (123)\\ &= -\max(0, (|w_j^{\mathrm{LS}}| - \lambda)) && (124)\\ &= \mathrm{sign}(w_j^{\mathrm{LS}}) \max(0, |w_j^{\mathrm{LS}}| - \lambda). && (125) \end{aligned}$$

In Eq. (122), we constrain w_j^{LR} to be less than or equal to zero (remember the assumptions in Case II we make) by adding a min operation. In Eq. (123), since $w_j^{\mathrm{LS}} < 0$, $-w_j^{\mathrm{LS}} = |w_j^{\mathrm{LS}}|$. In Eq. (124), we extract a minus sign out of the min operation and turn it into a max operation. In general, $q^\star = \min_\theta \ell(\theta) = -\max_\theta -\ell(\theta)$ for an arbitrary function. In Eq. (125), since $w_j^{\mathrm{LS}} < 0$, $\mathrm{sign}(w_j^{\mathrm{LS}}) = -1$.

In both cases, we get the desired form, and the corresponding vectorized expression is

$$w^{\mathrm{LR}} = \mathrm{sign}(w^{\mathrm{LS}}) \max(0, |w^{\mathrm{LS}}| - \lambda). \tag{126}$$

□

6 Lecture 6

6.1 On Primal Form of SVR with Soft Margin

Proof. The primal form of SVR with soft margin is

$$\begin{aligned} \min_{w,b,\xi,\xi^*} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*)\\ \text{s.t.} \quad & y^{(i)} - w^T x^{(i)} - b \le \epsilon + \xi_i,\ i = 1, \ldots, M,\\ & w^T x^{(i)} + b - y^{(i)} \le \epsilon + \xi_i^*,\ i = 1, \ldots, M,\\ & \xi_i, \xi_i^* \ge 0,\ i = 1, \ldots, M, \end{aligned} \tag{127}$$

from which we have

$$\begin{cases} y^{(i)} - w^T x^{(i)} - b \le \epsilon + \xi_i\\ w^T x^{(i)} + b - y^{(i)} \le \epsilon + \xi_i^* \end{cases} \quad \Longrightarrow \quad -(\epsilon + \xi_i^*) \le y^{(i)} - w^T x^{(i)} - b \le \epsilon + \xi_i. \tag{128}$$

That is,

$$\begin{aligned} |y^{(i)} - w^T x^{(i)} - b| &\le \max\{\epsilon + \xi_i, \epsilon + \xi_i^*\}\\ |y^{(i)} - w^T x^{(i)} - b| - \epsilon &\le \max\{\xi_i, \xi_i^*\} = \xi_i + \xi_i^*, \end{aligned} \tag{129}$$

where the last equation arises from the fact that ξ and ξ_i^* cannot be both strictly positive (*i.e.*, at least one of them must be zero). This is not hard to understand because we cannot simply have a prediction error, $y^{(i)} - w^T x^{(i)} - b$, that is both positive (exceeding ϵ with a positive ξ_i as penalty) and negative (exceeding $-\epsilon$ with a positive ξ_i^* as penalty). Note also that

$$\begin{cases} |y^{(i)} - w^T x^{(i)} - b| - \epsilon \leq \xi_i + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \implies \max\{0, |y^{(i)} - w^T x^{(i)} - b| - \epsilon\} \leq \xi_i + \xi_i^*. \tag{130}$$

$$\begin{aligned} \ell(w, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*) &= \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) + \sum_{i=1}^{M}\alpha_i[y^{(i)} - w^T x^{(i)} - b - \epsilon - \xi_i] \\ &+ \sum_{i=1}^{M}\alpha_i^*[w^T x^{(i)} + b - y^{(i)} - \epsilon - \xi_i^*] - \sum_{i=1}^{M}\beta_i\xi_i - \sum_{i=1}^{M}\beta_i^*\xi_i^* \\ &= \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{M}(C - \alpha_i - \beta_i)\xi_i + \sum_{i=1}^{M}(C - \alpha_i^* - \beta_i^*)\xi_i^* \\ &+ \sum_{i=1}^{M} y^{(i)}(\alpha_i - \alpha_i^*) - w^T\left(\sum_{i=1}^{M}(\alpha_i - \alpha_i^*)x^{(i)}\right) \\ &- b\sum_{i=1}^{M}(\alpha_i - \alpha_i^*) - \epsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*). \end{aligned} \tag{131}$$

Take the partial derivatives of ℓ w.r.t. w, b, ξ, and ξ^* and set them to zero:

$$\nabla_w \ell = w - \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)x^{(i)} = 0, \tag{132}$$

$$\nabla_b \ell = \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0 \tag{133}$$

$$\nabla_{\xi_i} \ell = C - \alpha_i - \beta_i = 0, \quad i = 1, \ldots, M, \tag{134}$$

$$\nabla_{\xi_i^*} \ell = C - \alpha_i^* - \beta_i^* = 0, \quad i = 1, \ldots, M. \tag{135}$$

Plugging Eq. (132), Eq. (134), Eq. (135), and Eq. (133) back into the Lagrangian (Eq. (131)), we obtain

$$g(\alpha, \alpha^*) = -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x^{(i)})^T x^{(j)} - \epsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} y^{(i)}(\alpha_i - \alpha_i^*) \tag{136}$$

Putting $g(\alpha, \alpha^*)$ with the constraints, we obtain the dual problem:

$$\begin{aligned}
\max_{\alpha,\alpha^*} \quad & -\frac{1}{2}\sum_{i,j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x^{(i)})^T x^{(j)} \\
& - \epsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} y^{(i)}(\alpha_i - \alpha_i^*), \\
\text{s.t.} \quad & \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0, \\
& C - \alpha_i - \beta_i = 0, \quad i = 1, \ldots, M, \\
& C - \alpha_i^* - \beta_i^* = 0, \quad i = 1, \ldots, M, \\
& \alpha_i, \beta_i, \alpha_i^*, \beta_i^* \geq 0, \quad i = 1, \ldots, M.
\end{aligned} \tag{137}$$

It is straightforward to show that

$$\begin{cases} C - \alpha_i - \beta_i = 0 \\ \alpha_i \geq 0 \\ \beta_i \geq 0 \end{cases} \quad \Longleftrightarrow \quad \begin{cases} \alpha_i \geq 0 \\ \alpha_i \leq C. \end{cases} \tag{138}$$

Similarly,

$$\begin{cases} C - \alpha_i^* - \beta_i^* = 0 \\ \alpha_i^* \geq 0 \\ \beta_i^* \geq 0 \end{cases} \quad \Longleftrightarrow \quad \begin{cases} \alpha_i^* \geq 0 \\ \alpha_i^* \leq C. \end{cases} \tag{139}$$

Finally, the dual problem of SVR is

$$\begin{aligned}
\max_{\alpha,\alpha^*} \quad & -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x^{(i)})^T x^{(j)} \\
& - \epsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} y^{(i)}(\alpha_i - \alpha_i^*), \\
\text{s.t.} \quad & \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0, \\
& 0 \leq \alpha_i, \alpha_i^* \leq C, \; i = 1, \ldots, M.
\end{aligned} \tag{140}$$

□

6.2 Dual Form of Ridge Regression

Proof. Recall the objective of ridge regression

$$\min_{w} \quad \frac{1}{2}\|y - Xw\|_2^2 + \frac{\alpha}{2}\|w\|_2^2, \tag{141}$$

which is equivalent to the following optimization problem:

$$\begin{aligned} \min_{w,\xi} \quad & \frac{1}{2}\|\xi\|_2^2 + \frac{\alpha}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y - Xw = \xi, \end{aligned} \tag{142}$$

where we add a dummy vector $\xi \in \mathbb{R}^{M\times 1}$. We now form the Lagrangian:

$$\ell(w,\xi,\beta) = \frac{1}{2}\|\xi\|_2^2 + \frac{\alpha}{2}\|w\|_2^2 + \beta^T\left(y - Xw - \xi\right), \tag{143}$$

where $\beta \in \mathbb{R}^{M\times 1}$ is the dual vector containing M Lagrange multipliers. Taking the partial derivatives of ℓ w.r.t. primal variables w and ξ, and setting them to zero, we have

$$\nabla_w \ell = \alpha w - X^T\beta = 0, \tag{144}$$

$$\nabla_\xi \ell = \xi - \beta = 0. \tag{145}$$

Plugging Eq. (144) and Eq. (145) back into the Lagrangian, we obtain the Lagrange dual function:

$$\begin{aligned} g(\beta) &= \frac{1}{2}\beta^T\beta + \frac{\alpha}{2}\left\|\frac{X^T\beta}{\alpha}\right\|_2^2 + \beta^T\left(y - X\frac{X^T\beta}{\alpha} - \beta\right) \\ &= -\frac{1}{2}\beta^T\beta + \frac{\alpha}{2}\left(\frac{X^T\beta}{\alpha}\right)^T\frac{X^T\beta}{\alpha} + \beta^T y - \frac{\beta^T XX^T\beta}{\alpha} \\ &= -\frac{\beta^T(\alpha I_M)\beta}{2\alpha} - \frac{\beta^T XX^T\beta}{2\alpha} + \beta^T y \\ &= -\frac{\beta^T(XX^T + \alpha I_M)\beta}{2\alpha} + \beta^T y. \end{aligned} \tag{146}$$

Maximizing g w.r.t. β, we have

$$\nabla_\beta g = -\frac{(XX^T + \alpha I_M)\beta}{\alpha} + y = 0 \tag{147}$$

$$\beta = \alpha(XX^T + \alpha I_M)^{-1}y. \tag{148}$$

Combining Eq. (144) and Eq. (148), we have

$$w = \frac{X^T\beta}{\alpha} = X^T(XX^T + \alpha I_M)^{-1}y, \tag{149}$$

as desired. □

6.3 Bell Number

In combinatorial mathematics, the Bell number[14] counts the possible partitions of a set. For concreteness, let's suppose that we are partitioning the set $\mathcal{D} = \{x^{(1)}, x^{(2)}, \ldots, x^{(M+1)}\}$,

[14] https://en.wikipedia.org/wiki/Bell_number

which consists of $M+1$ elements (*i.e.*, data cases). Focus first on the cluster containing the element $x^{(1)}$. Let k denote the number of elements other than $x^{(1)}$ that belong to this cluster. We can choose these elements in $\binom{M}{k}$ ways[15]. Having formed this cluster, we partition the remaining $M+1-(k+1)=M-k$ elements in B_{M-k} ways. Summing over k from 0 to M (as we have already picked out $x^{(1)}$, the maximum value of k is M rather than $M+1$) gives

$$B_{M+1}=\sum_{k=0}^{M}\binom{M}{k}B_{M-k}. \tag{150}$$

By the symmetry of the binomial coefficients, *i.e.*, $\binom{M}{k}=\binom{M}{M-k}$, this expression is equivalent to

$$B_{M+1}=\sum_{k=0}^{M}\binom{M}{k}B_{M-k}=\sum_{k=0}^{M}\binom{M}{M-k}B_{M-k}=\sum_{k=0}^{M}\binom{M}{k}B_{k}, \tag{151}$$

where $B_0=1$ and $B_1=1$ (because there is exactly one partition of the empty set and the set containing only one element).

7 Lecture 7

7.1 Coordinate Ascent Interpretation of the EM Algorithm

Define the following objective function

$$J(q,\theta)=\sum_{i=1}^{M}\sum_{z^{(i)}=1}^{K}q(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{q(z^{(i)})}. \tag{152}$$

It can be shown that EM can be viewed as coordinate ascent on J, in which the E-step maximizes it with respect to q and the M-step maximizes it with respect to θ.

In the E-step, we must take account of the constraint

$$\sum_{z^{(i)}=1}^{K}q(z^{(i)})=1. \tag{153}$$

This can be achieved using the method of Lagrange multipliers by maximizing

$$\bar{J}(q,\theta)=\sum_{i=1}^{M}\sum_{z^{(i)}=1}^{K}q(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{q(z^{(i)})}+\lambda\left(1-\sum_{z^{(i)}=1}^{K}q(z^{(i)})\right). \tag{154}$$

Taking the partial derivative of $\bar{J}$ w.r.t. $q(z^{(i)})$ and setting it to zero, we have

$$\begin{aligned}
&\log\frac{p(x^{(i)},z^{(i)};\theta)}{q(z^{(i)})}+q(z^{(i)})\frac{q(z^{(i)})}{p(x^{(i)},z^{(i)};\theta)}\left(-\frac{p(x^{(i)},z^{(i)};\theta)}{q(z^{(i)})^2}\right)-\lambda=0 &(155)\\
&\log\frac{p(x^{(i)},z^{(i)};\theta)}{q(z^{(i)})}-1-\lambda=0\\
&q(z^{(i)})=\frac{p(x^{(i)},z^{(i)};\theta)}{e^{1+\lambda}}. &(156)
\end{aligned}$$

[15] https://en.wikipedia.org/wiki/Combination

In Eq. (155), we use the chain rule of the derivatives[16] of the composition of two or more functions, *i.e.*,

$$\frac{\partial f(\theta)g(\theta)}{\partial\theta_j} = f(\theta)\frac{\partial g(\theta)}{\partial\theta_j} + g(\theta)\frac{\partial f(\theta)}{\partial\theta_j} \tag{157}$$

and

$$\frac{\partial f(g(\theta))}{\partial\theta_j} = \frac{df(g(\theta))}{dg(\theta)}\frac{\partial g(\theta)}{\partial\theta_j}. \tag{158}$$

Since

$$\sum_{z^{(i)}=1}^{K} q(z^{(i)}) = \frac{\sum_{z^{(i)}=1}^{K} p(x^{(i)}, z^{(i)}; \theta)}{e^{1+\lambda}} = \frac{p(x^{(i)}; \theta)}{e^{1+\lambda}} = 1, \tag{159}$$

we have

$$q(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)}; \theta)}{e^{1+\lambda}} = \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} = p(z^{(i)}|x^{(i)}; \theta), \tag{160}$$

as desired. In the M-step, we maximize J w.r.t. the parameter vector θ while keeping q fixed

$$\arg\max_{\theta} J(q, \theta) = \arg\max_{\theta} \sum_{i=1}^{M} \sum_{z^{(i)}=1}^{K} q(z^{(i)}) \log p(x^{(i)}, z^{(i)}; \theta). \tag{161}$$

7.2 Gaussian Mixture Models Revisited[17]

Armed with the EM algorithm, let's go back to our old example of fitting the parameters $\{\pi, \mu, \Sigma\}$ in Gaussian mixture models.

The E-step is easy. Following our EM algorithm derivation, we simply calculate

$$\begin{aligned} z_j^{(i)} = q(z^{(i)} = j) = p(z^{(i)} = j|x^{(i)}) &= \frac{p(x^{(i)}, z^{(i)} = j)}{p(x^{(i)})} \\ &= \frac{p(x^{(i)}|z^{(i)} = j)p(z^{(i)} = j)}{\sum_{k=1}^{K} p(x^{(i)}|z^{(i)} = k)p(z^{(i)} = k)} \\ &= \frac{\pi_j \mathcal{N}(x^{(i)}; \mu_j, \Sigma_j)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}; \mu_k, \Sigma_k)}, \end{aligned} \tag{162}$$

where $\{\pi_j, \mu_j, \Sigma_j\}_{j=1}^{K}$ are obtained from the previous M-step.

[16] `https://en.wikipedia.org/wiki/Chain_rule`

[17] Inspired by Andrew Ng's lecture notes.

Next, in the M-step, we need to maximize, with respect to our parameters $\{\pi, \mu, \Sigma\}$,

$$\begin{aligned}
&\sum_{i=1}^{M}\sum_{z^{(i)}=1}^{K} q(z^{(i)}) \log p(x^{(i)}, z^{(i)}; \pi, \mu, \Sigma) \\
&\quad = \sum_{i=1}^{M}\sum_{j=1}^{K} q(z^{(i)} = j) \log \left(p(x^{(i)}|z^{(i)} = j; \mu_j, \Sigma_j) p(z^{(i)} = j; \pi_j)\right) \\
&\quad = \sum_{i=1}^{M}\sum_{j=1}^{K} z_j^{(i)} \log \left(\frac{1}{(2\pi)^{N/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \cdot \pi_j \right). \qquad (163)
\end{aligned}$$

Let's first derive the M-step update for the parameters $\{\pi_j\}_{j=1}^K$. Grouping together the terms that depend only on $\{\pi_j\}_{j=1}^K$ in Eq. (163), we find that we need to maximize

$$\sum_{i=1}^{M}\sum_{j=1}^{K} z_j^{(i)} \log \pi_j. \qquad (164)$$

However, there is an additional constraint that the π_j's should sum up to one, since they represent the probability $\pi_j = p(z^{(i)} = j; \pi_j)$. To deal with the constraint that $\sum_{j=1}^K \pi_j = 1$, we construct the Lagrangian

$$\sum_{i=1}^{M}\sum_{j=1}^{K} z_j^{(i)} \log \pi_j + \lambda \left(1 - \sum_{j=1}^{K} \pi_j\right), \qquad (165)$$

where λ is the Lagrange multiplier. Taking the derivative w.r.t. π_j and setting it to zero, we find

$$\begin{aligned}
\sum_{i=1}^{M} \frac{z_j^{(i)}}{\pi_j} - \lambda &= 0 \\
\pi_j &= \frac{\sum_{i=1}^{M} z_j^{(i)}}{\lambda}. \qquad (166)
\end{aligned}$$

Using the constraint that $\sum_{j=1}^K \pi_j = 1$, we easily find that

$$\lambda = \sum_{j=1}^{K}\sum_{i=1}^{M} z_j^{(i)} = \sum_{i=1}^{M}\sum_{j=1}^{K} z_j^{(i)} = \sum_{i=1}^{M} 1 = M. \qquad (167)$$

We therefore have our M-step updates for the parameter π_j:

$$\pi_j = \frac{\sum_{i=1}^{M} z_j^{(i)}}{M}. \qquad (168)$$

Let's then maximize the objective in Eq. (163) with respect to μ_j. If we take the derivative w.r.t. μ_j, we find

$$\begin{aligned}
&\nabla_{\mu_j} \sum_{i=1}^{M} \sum_{j=1}^{K} z_j^{(i)} \log \left(\frac{1}{(2\pi)^{N/2} |\Sigma_j|^{1/2}} \exp \left(-\frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \cdot \pi_j \right) \\
&= -\frac{1}{2} \nabla_{\mu_j} \sum_{i=1}^{M} z_j^{(i)} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \\
&= -\frac{1}{2} \sum_{i=1}^{M} z_j^{(i)} \nabla_{\mu_j} \left((x^{(i)})^T \Sigma_j^{-1} x^{(i)} - (x^{(i)})^T \Sigma_j^{-1} \mu_j - \mu_j^T \Sigma_j^{-1} x^{(i)} + \mu_j^T \Sigma_j^{-1} \mu_j \right) \\
&= -\frac{1}{2} \sum_{i=1}^{M} z_j^{(i)} \nabla_{\mu_j} (-2 (x^{(i)})^T \Sigma_j^{-1} \mu_j + \mu_j^T \Sigma_j^{-1} \mu_j) \\
&= \sum_{i=1}^{M} z_j^{(i)} (\Sigma_j^{-1} x^{(i)} - \Sigma_j^{-1} \mu_j). \qquad (169)
\end{aligned}$$

Setting Eq. (169) to zero and solving for μ_j therefore yields the update rule

$$\mu_j = \frac{\sum_{i=1}^{M} z_j^{(i)} x^{(i)}}{\sum_{i=1}^{M} z_j^{(i)}}, \qquad (170)$$

which was what we had in our lecture slides. We last carry out the derivations of the update rule for Σ_j. If we take the derivative of the objective in Eq. (163) w.r.t. Σ_j, we find

$$\begin{aligned}
&\nabla_{\Sigma_j} \sum_{i=1}^{M} \sum_{j=1}^{K} z_j^{(i)} \log \left(\frac{1}{(2\pi)^{N/2} |\Sigma_j|^{1/2}} \exp \left(-\frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \cdot \pi_j \right) \\
&= -\frac{1}{2} \sum_{i=1}^{M} z_j^{(i)} \nabla_{\Sigma_j} \left(\log |\Sigma_j| + (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \\
&= -\frac{1}{2} \sum_{i=1}^{M} z_j^{(i)} \left(\Sigma_j^{-1} - \Sigma_j^{-1} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T \Sigma_j^{-1} \right). \qquad (171)
\end{aligned}$$

In Eq. (171), we have made use of two derivatives of matrices[18]

$$\frac{\partial \log |X|}{\partial X} = (X^{-1})^T = X^{-1}, \qquad (172)$$

and

$$\frac{\partial a^T X^{-1} b}{\partial X} = -(X^{-1})^T a b^T (X^{-1})^T = -X^{-1} a b^T X^{-1}, \qquad (173)$$

[18] Please refer to the Eq. (49) and Eq. (61) in The Matrix Cookbook.

for a symmetric matrix X. Setting Eq. (171) to zero, we have

$$\begin{aligned}
&\sum_{i=1}^{M} z_j^{(i)} \left(\Sigma_j^{-1} - \Sigma_j^{-1}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T \Sigma_j^{-1}\right) = 0 \\
&\Sigma_j^{-1} \left(\sum_{i=1}^{M} z_j^{(i)} I - \sum_{i=1}^{M} z_j^{(i)}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T \Sigma_j^{-1}\right) = 0 \\
&\sum_{i=1}^{M} z_j^{(i)} I - \sum_{i=1}^{M} z_j^{(i)}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T \Sigma_j^{-1} = 0 \qquad (174) \\
&\Sigma_j = \frac{\sum_{i=1}^{M} z_j^{(i)}(x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{M} z_j^{(i)}}, \qquad (175)
\end{aligned}$$

as desired. In Eq. (174), we exploit the fact that the null space of Σ_j^{-1} is $\{0\}$.

7.3 Closed-Form Updates of Alternating Least Squares

We first recall the definition of the *trace* of a square matrix $A \in \mathbb{R}^{N \times N}$

$$\text{tr}A = \sum_{i=1}^{N} A_{ii} = A_{11} + A_{22} + \cdots + A_{NN}. \qquad (176)$$

The trace is a linear mapping. That is,

$$\text{tr}(A + B) = \text{tr}A + \text{tr}B, \qquad (177)$$

and

$$\text{tr}cA = c\text{tr}A. \qquad (178)$$

We note a trivial but useful property of the trace

$$\text{tr}A = \text{tr}A^T = \sum_{i=1}^{N} A_{ii}. \qquad (179)$$

This result implies that for $A \in \mathbb{R}^{M \times N}$ and $B \in \mathbb{R}^{M \times N}$

$$\text{tr}AB^T = \text{tr}(AB^T)^T = \text{tr}BA^T \qquad (180)$$

and

$$\text{tr}A^T B = \text{tr}(A^T B)^T = \text{tr}B^T A. \qquad (181)$$

We prove another important property of the trace: for $A \in \mathbb{R}^{M \times N}$ and $B \in \mathbb{R}^{N \times M}$, $\text{tr}AB = \text{tr}BA$.

Proof.

$$\mathrm{tr}AB = \sum_{i=1}^{M}(AB)_{ii} = \sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}b_{ji} = \sum_{j=1}^{N}\sum_{i=1}^{M} b_{ji}a_{ij} = \sum_{j=1}^{N}(BA)_{jj} = \mathrm{tr}BA. \tag{182}$$

□

More generally, the trace is invariant under cyclic permutations, which can be easily proved using Eq. (182). For example, for $A \in \mathbb{R}^{M\times N}$, $B \in \mathbb{R}^{N\times L}$, and $C \in \mathbb{R}^{L\times M}$, we have

$$\mathrm{tr}ABC = \mathrm{tr}A(BC) = \mathrm{tr}BCA = \mathrm{tr}B(CA) = \mathrm{tr}CAB. \tag{183}$$

We now prove that for $A \in \mathbb{R}^{M\times N}$

$$\|A\|_F^2 = \mathrm{tr}AA^T = \mathrm{tr}A^TA. \tag{184}$$

Proof.

$$\begin{aligned}\|A\|_F^2 &= \sum_{i=1}^{M}\sum_{j=1}^{N} A_{ij}^2 = \sum_{i=1}^{M}\left(\sum_{j=1}^{N} A_{ij}A_{ij}\right) = \sum_{i=1}^{M}\left(\sum_{j=1}^{N} A_{ij}A_{ji}^T\right) &(185)\\ &= \sum_{i=1}^{M}(AA^T)_{ii} = \mathrm{tr}AA^T. &(186)\end{aligned}$$

In Eq. (185), we note the fact that the entry in the ith row and jth column of A is equal to the entry in the jth row and ith column of A^T. □

Using Eq. (184), we are able to decompose the loss function in linear dimensionality reduction as

$$\begin{aligned}\ell(Z,B) = \|X - ZB\|_F^2 &= \mathrm{tr}(X - ZB)(X - ZB)^T\\ &= \mathrm{tr}(XX^T - X(ZB)^T - ZBX^T + ZB(ZB)^T)\\ &= \mathrm{tr}(XX^T - 2ZBX^T + ZBB^TZ^T) &(187)\\ &= \mathrm{tr}XX^T - 2\mathrm{tr}ZBX^T + \mathrm{tr}ZBB^TZ^T. &(188)\end{aligned}$$

In Eq. (187), we make use of the trace identity $\mathrm{tr}AB^T = \mathrm{tr}BA^T$. To derive the closed-form update rules of alternating least squares in solving the linear dimensionality reduction problem, we first take the derivative of ℓ w.r.t. Z

$$\begin{aligned}\frac{\partial \ell(Z,B)}{\partial Z} &= -2\frac{\partial}{\partial Z}\mathrm{tr}Z(BX^T) + \frac{\partial}{\partial Z}\mathrm{tr}Z(BB^T)Z^T\\ &= -2XB^T + ZBB^T + Z(BB^T)^T\\ &= -2XB^T + 2ZBB^T, &(189)\end{aligned}$$

where we make use of two derivatives of traces[19]

$$\frac{\partial}{\partial X}\mathrm{tr}XA = A^T \tag{190}$$

[19] Please refer to the Eq. (100) and Eq. (111) in the The Matrix Cookbook.

and

$$\frac{\partial}{\partial X}\mathrm{tr}XAX^T = XA + XA^T. \tag{191}$$

Setting the partial derivative in Eq. (189) to zero, we have

$$\begin{aligned} ZBB^T &= XB^T && (192)\\ Z &= XB^T(BB^T)^{-1} && (193)\\ Z^T &= (XB^T(BB^T)^{-1})^T && (194)\\ &= (BB^T)^{-1}BX^T. && (195) \end{aligned}$$

In Eq. (195), we exploit the fact that the inverse of the transpose of an invertible matrix is equal to the transpose of its inverse, *i.e.*, $(A^T)^{-1} = (A^{-1})^T$. Similarly, taking the derivative of ℓ w.r.t. B, we have

$$\begin{aligned} \frac{\partial \ell(Z,B)}{\partial B} &= -2\frac{\partial}{\partial B}\mathrm{tr}ZBX^T + \frac{\partial}{\partial B}\mathrm{tr}ZBB^TZ^T\\ &= -2\frac{\partial}{\partial B}\mathrm{tr}B(X^TZ) + \frac{\partial}{\partial B}\mathrm{tr}(Z^TZ)BB^T\\ &= -2Z^TX + Z^TZB + (Z^TZ)^TB\\ &= -2Z^TX + 2Z^TZB, && (196) \end{aligned}$$

where we use the derivatives of trace[20]

$$\frac{\partial}{\partial X}\mathrm{tr}AXX^T = AX + A^TX. \tag{197}$$

Setting the partial derivative in Eq. (196) to zero, we have

$$\begin{aligned} Z^TZB &= Z^TX\\ B &= (Z^TZ)^{-1}Z^TX. && (198) \end{aligned}$$

7.4 Proof of Eckart-Young-Mirsky Theorem

Proof. We first prove that the Frobenius norm is preserved under orthogonal transformations. That is,

$$\|UX\|_F^2 = \|XU\|_F^2 = \|X\|_F^2, \tag{199}$$

where U is an orthogonal matrix, *i.e.*, $U^TU = UU^T = I$. Using Eq. (184), we have

$$\|UX\|_F^2 = \mathrm{tr}(UX)^TUX = \mathrm{tr}X^TU^TUX = \mathrm{tr}X^TX = \|X\|_F^2. \tag{200}$$

Similarly,

$$\|XU\|_F^2 = \mathrm{tr}XU(XU)^T = \mathrm{tr}XUU^TX^T = \mathrm{tr}XX^T = \|X\|_F^2. \tag{201}$$

[20] Please refer to the Eq. (109) in the The Matrix Cookbook.

Starting from the SVD of $X = U\Sigma_R V^T$, where $U^T U = UU^T = I$ and $V^T V = VV^T = I$, we have

$$\begin{aligned}\|X - Y\|_F^2 &= \|U\Sigma_R V^T - Y\|_F^2 \\ &= \|U^T U\Sigma_R V^T - U^T Y\|_F^2 \\ &= \|\Sigma_R V^T - U^T Y\|_F^2 \\ &= \|\Sigma_R V^T V - U^T Y V\|_F^2 \\ &= \|\Sigma_R - U^T Y V\|_F^2.\end{aligned}$$

Denoting $Z = U^T Y V$, an $M \times N$ matrix of rank K, a direct calculation gives

$$\|\Sigma - Z\|_F^2 = \sum_{ij} |\Sigma_{ij} - Z_{ij}|^2 = \sum_{i=1}^{R} |\sigma_i - Z_{ii}|^2 + \sum_{i>R} |Z_{ii}|^2 + \sum_{i\neq j} |Z_{ij}|^2,$$

which is minimal when all the non diagonal terms of Z equal to zero, and so are all Z_{ii} for $i > R$. This implies that Y shares the same left and right singular vectors with X, *i.e.*, $Y = UZV^T$ with Z being a diagonal matrix. Finally, the minimum of $\sum_{i=1}^{R} |\sigma_i - Z_{ii}|^2$ among all $Z_{11}, \ldots, Z_{RR}$ is attained when $Z_{ii} = \sigma_i$ for $i = 1, \ldots, K$ and all other Z_{ii} are zero, *i.e.*, $Z = \Sigma_K$[21]. Therefore, we have $Y^\star = U\Sigma_K V^T$ is the best rank K approximation of X under the Frobenius norm. □

8 Lecture 8

8.1 Linear Independence[22]

A set of vectors $\{v_1, v_2, \ldots, v_N\}$ is **linear independent** if the vector equation

$$a_1 v_1 + a_2 v_2 + \cdots + a_N v_N = 0 \tag{202}$$

has only the trivial solution $a_1 = a_2 = \cdots = a_N = 0$. Otherwise, the set $\{v_1, v_2, \ldots, v_N\}$ is **linearly dependent**. In other words, $\{v_1, v_2, \ldots, v_N\}$ is linearly dependent if there exist numbers $a_1, a_2, \ldots, a_N$ not all equal to zero, such that

$$a_1 v_1 + a_2 v_2 + \cdots + a_N v_N = 0. \tag{203}$$

In this case, without loss of generality, we assume $a_1 \neq 0$ and we are able to write v_1 as a linear combination of the rest $N - 1$ vectors

$$v_1 = -\frac{a_2}{a_1} v_2 - \cdots - \frac{a_N}{a_1} v_N, \tag{204}$$

manifesting the linear dependence relation.

To enhance this extremely important concept in linear algebra, let's prove some facts about linear independence.

[21] Note that since the rank of Y is less than or equal to K, we can have at most K non-zero sigular values in Z.

[22] Inspired by Dan Margalit and Joseph Rabinoff

- Two vectors are linearly dependent if and only if they are collinear, *i.e.*, one is a scalar multiple of the other.

  Proof. If $v_1 = av_2$ then $v_1 - av_2 = 0$, so $\{v_1, v_2\}$ is linearly dependent (note that the coefficient associated with v_1 is not zero). In the other direction, if $\{v_1, v_2\}$ is linearly dependent, we have $a_1v_1 + a_2v_2 = 0$ with $a_1 \neq 0$ (without loss of generality), then $v_1 = -\frac{a_2}{a_1}v_2$. □

- Any set containing the zero vector is linearly dependent.

  Proof. It is easy to produce a linear dependence relation if one vector is the zero vector: for instance, if $v_1 = 0$, then

  $$1v_1 + 0v_2 + \cdots + 0v_N = 0. \tag{205}$$

  Therefore, the set $\{v_1, v_2, \ldots, v_N\}$ is linearly dependent. □

- If a subset of $\{v_1, v_2, \ldots, v_N\}$ is linearly dependent, then $\{v_1, v_2, \ldots, v_N\}$ is linearly dependent as well.

  Proof. Suppose that $\{v_1, v_2, \ldots, v_K\}$ is linear dependent with $K < N$. This means that there is an equation of linear dependence

  $$a_1v_1 + a_2v_2 + \cdots + a_Kv_K = 0, \tag{206}$$

  with at least one of $a_1, a_2, \ldots, a_K$ nonzero. There is also an equation of linear dependence among $\{v_1, v_2, \ldots, v_N\}$ since we can take the coefficients of $v_{K+1}, \ldots, v_N$ to all be zero:

  $$a_1v_1 + a_2v_2 + \cdots + a_Kv_K + 0v_{K+1} + \cdots + 0v_N = 0. \tag{207}$$

  □

- If the set of non-zero vectors $\{v_1, v_2, \ldots, v_N\}$ are orthogonal to each other, *i.e.*, $v_i^T v_j = 0$ for $i \neq j$, then the set $\{v_1, v_2, \ldots, v_N\}$ is linearly independent.

  Proof. Consider the linear combination

  $$a_1v_1 + a_2v_2 + \cdots + a_Nv_N = 0. \tag{208}$$

  Our goal is to show that the only solution to the above equation is $a_1 = a_2 = \cdots = a_N = 0$. To achieve this, we left multiply v_i^T to obtain

  $$a_1v_i^Tv_1 + a_2v_i^Tv_2 + \cdots + a_Nv_i^Tv_N = 0, \quad i \in \{1, 2, \ldots, N\}. \tag{209}$$

  Due to the orthogonality of v_i's, all the terms but the ith one is zero, and hence we have

  $$a_iv_i^Tv_i = a_i\|v_i\|_2^2 = 0, \quad i \in \{1, 2, \ldots, N\}. \tag{210}$$

  Since v_i is a non-zero vector, its length $\|v_i\|_2$ is non-zero. It follows that

  $$a_i = 0, \quad i \in \{1, 2, \ldots, N\}. \tag{211}$$

  □

8.2 Look at PCA via Two Lens

In the lecture slides, we have derived PCA by looking for the directions of maximum variance. Here we show that PCA can also be derived by picking the directions that minimize the approximation error arising from projecting the data onto the K-dimensional subspace spanned by them.

Proof. Without loss of generality, we assume $K = 1$. The approximation error, *i.e.*, the mean squared error, between $\{x^{(i)}\}$ and their projections are

$$\begin{aligned}\frac{1}{M}\sum_{i=1}^{M}\|x^{(i)} - (v^T x^{(i)})v\|_2^2 &= \frac{1}{M}\sum_{i=1}^{M}\left(\|x^{(i)}\|_2^2 + (v^T x^{(i)})^2\|v\|_2^2 - 2(v^T x^{(i)})^2\right)\\ &= \frac{1}{M}\sum_{i=1}^{M}\left(\|x^{(i)}\|_2^2 - (v^T x^{(i)})^2\right), \qquad (212)\end{aligned}$$

where $\|v\|_2^2 = 1$. Since all $\{x^{(i)}\}$ are given and therefore $\|x^{(i)}\|_2^2$ are fixed for all i, minimizing the approximation error is equivalent to maximizing the variance of the projections

$$\arg\min_v \sum_{i=1}^{M}\left(\|x^{(i)}\|_2^2 - (v^T x^{(i)})^2\right) = \arg\max_v \frac{1}{M}\sum_{i=1}^{M}(v^T x^{(i)})^2 = \arg\max_v v^T\Sigma v. \qquad (213)$$

□

8.3 Last Ingredient of Kernel PCA

In lecture slides, we derived kernel PCA with an implicit assumption: the data points, either in the original space $\{x\}$ or in the high-dimensional feature space $\{\phi(x)\}$, have zero mean. Therefore, the covariance of two data points is the same as their correlation. While the data points can be easily de-meaned in the original space by subtracting the mean vector from all points, it is done differently in the feature space. The de-meaned data points in the feature space is:

$$\tilde{\phi}(x^{(i)}) = \phi(x^{(i)}) - \frac{1}{M}\sum_{k=1}^{M}\phi(x^{(k)}). \qquad (214)$$

But this cannot be actually carried out as the mapping is not explicit and $\phi(x^{(i)})$ is never available. Fortunately, we can still obtain the kernel matrix $\tilde{K}$ for the zero-mean data points $\tilde{\phi}(x)$ in terms of K for $\phi(x)$, where $K_{ij} = \phi(x^{(i)})^T\phi(x^{(j)})$. Specifically,

$$\begin{aligned}\tilde{K}_{ij} = \tilde{\phi}(x^{(i)})^T\tilde{\phi}(x^{(j)}) &= \left(\phi(x^{(i)}) - \frac{1}{M}\sum_{k=1}^{M}\phi(x^{(k)})\right)^T\left(\phi(x^{(j)}) - \frac{1}{M}\sum_{k=1}^{M}\phi(x^{(k)})\right)\\ &= \phi(x^{(i)})^T\phi(x^{(j)}) - \frac{1}{M}\sum_{k=1}^{M}\phi(x^{(i)})^T\phi(x^{(k)}) - \frac{1}{M}\sum_{k=1}^{M}\phi(x^{(k)})^T\phi(x^{(j)}) + \frac{1}{M^2}\sum_{k=1}^{M}\sum_{k'=1}^{M}\phi(x^{(k)})^T\phi(x^{(k')})\\ &= K_{ij} - K_{i\bullet}1_M - 1_M^T K_{\bullet j} + 1_M^T K 1_M, \qquad (215)\end{aligned}$$

where $K_{i\bullet}$ denotes the ith row of K, $K_{\bullet j}$ denotes the jth column of K, and 1_M is an $M \times 1$ column vector with each entry equal to $1/M$. The corresponding matrix equation is

$$\begin{aligned}\tilde{K} &= K - K1_{M\times M} - 1_{M\times M}K + 1_{M\times M}K1_{M\times M} \\ &= (I - 1_{M\times M})K(I - 1_{M\times M}),\end{aligned} \tag{216}$$

where $1_{M\times M}$ is an $M \times M$ matrix with each entry equal to $1/M$.

9 Lecture 9

9.1 Justification of the Update Rule of Perceptron

Proof. The perceptron algorithm is mistake-driven. Suppose the input is an $N+1$ dimensional vector $x \in \mathbb{R}^{N+1}$ and the output is a binary label $y \in \{-1, 1\}$, respectively. The perceptron algorithm makes prediction using

$$\hat{y} = \begin{cases} 1, & \text{if } w^T x \geq 0 \\ -1 & \text{if } w^T x < 0, \end{cases} \tag{217}$$

where $w \in \mathbb{R}^{N+1}$ is a learnable weight vector. Let $w^{(t)}$ be the weights used when it makes the t-th mistake. If the example is positive, *i.e.*, $y = 1$, we have $x^T w^{(t)} < 0$. We update $w^{(t)}$ by

$$w^{(t+1)} = w^{(t)} + \alpha y x = w^{(t)} + \alpha x, \tag{218}$$

where α is a positive update step, *i.e.*, the learning rate.

Now the inner product becomes

$$x^T w^{(t+1)} = x^T(w^{(t)} + \alpha x) = x^T w^{(t)} + \alpha \|x\|_2^2 > x^T w^{(t)}. \tag{219}$$

That is, the perceptron update will increase the score for a misclassified positive example. Similarly, if the example is negative, *i.e.*, $y = -1$, we have $x^T w^{(t)} \geq 0$. We update $w^{(t)}$ by

$$w^{(t+1)} = w^{(t)} - \alpha x, \tag{220}$$

and the inner product will be

$$x^T w^{(t+1)} = x^T(w^{(t)} - \alpha x) = x^T w^{(t)} - \alpha \|x\|_2^2 < x^T w^{(t)}. \tag{221}$$

In other words, the perceptron update will decrease the score for a misclassified negative example. Altogether, we have justified the update rule of the perceptron algorithm for a misclassified example

$$w^{(t+1)} = w^{(t)} + \alpha y x. \tag{222}$$

□

9.2 Proof of the Perceptron Theorem[23]

Let the training set be $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, M\}$. Assume that $\|x^{(i)}\|_2 \leq a$ for all i,. which can be achieved via feature normalization. Assume further that there exists a unit-length vector u ($\|u\|_2 = 1$) such that $y^{(i)}(u^T x^{(i)}) \geq b$ for all examples in $\mathcal{D}$. This implies the training set $\mathcal{D}$ is linear separable with geometric margin[24] at least b. Then the total number of mistakes that the perceptron algorithm needs to correct until convergence is at most $(a/b)^2$.

Proof. The perceptron updates its weights only using those examples on which it makes mistakes. Let $w^{(t)}$ be the weights that were being used when it made its t-th mistake. So, $w^{(1)} = 0$ (since the weights are initialized to zero), and if the t-th mistake was on the example $(x^{(i)}, y^{(i)})$, then $y^{(i)} \neq \hat{y}^{(i)}$, which implies that

$$y^{(i)}(x^{(i)})^T w^{(t)} \leq 0. \tag{223}$$

Also, from the perceptron learning rule, we would have that

$$w^{(t+1)} = w^{(t)} + y^{(i)} x^{(i)}, \tag{224}$$

assuming the learning rate $\alpha = 1$ without loss of generality. We then have

$$\begin{aligned} u^T w^{(t+1)} &= u^T(w^{(t)} + y^{(i)} x^{(i)}) \\ &\geq u^T w^{(t)} + b. \end{aligned} \tag{225}$$

A straightforward inductive argument implies that

$$u^T w^{(t+1)} \geq u^T w^{(1)} + tb = tb. \tag{226}$$

Also, we have that

$$\begin{aligned} \|w^{(t+1)}\|_2^2 &= \|w^{(t)} + y^{(i)} x^{(i)}\|_2^2 \\ &= \|w^{(t)}\|_2^2 + \|y^{(i)} x^{(i)}\|_2^2 + 2y^{(i)}(x^{(i)})^T w^{(t)} \\ &= \|w^{(t)}\|_2^2 + \|x^{(i)}\|_2^2 + 2y^{(i)}(x^{(i)})^T w^{(t)} && (227) \\ &\leq \|w^{(t)}\|_2^2 + \|x^{(i)}\|_2^2 && (228) \\ &\leq \|w^{(t)}\|_2^2 + a^2. && (229) \end{aligned}$$

In Eq. (227), we use $(y^{(i)})^2 = 1$. In Eq. (228), we refer to our assumption in Eq. (223). Moreover, again by applying a straightforward inductive argument, we see that

$$\|w^{(t+1)}\|_2^2 \leq \|w^{(1)}\|_2^2 + ta^2 = ta^2. \tag{230}$$

[23] Inspired by Andrew Ng's lecture notes.

[24] Recall the definition of geometric margin when we study SVM.

Putting Eq. (226) and Eq. (230) together, we find that

$$\begin{aligned}\sqrt{t}a &\geq \|w^{(t+1)}\|_2 \\ &= \|w^{(t+1)}\|_2\|u\|_2 \\ &\geq \|w^{(t+1)}\|_2\|u\|_2 \cos\theta, \quad \text{where } \theta \text{ is the angle between } w^{(t+1)} \text{ and } u \\ &= u^T w^{(t+1)} \\ &\geq tb. \end{aligned} \tag{231}$$

Our result implies that

$$t \leq \left(\frac{a}{b}\right)^2. \tag{232}$$

Hence, if the perceptron made a t-th mistake, t must be less than or equal to $(a/b)^2$. That is, the perceptron always converges as long as the data set $\mathcal{D}$ is linearly separable (*i.e.*, there exists a unit-length vector u such that $y^{(i)}(u^T x^{(i)}) \geq b$ for all $\{x^{(i)}\}$).

□

10 Lecture 12

10.1 On the Forward Diffusion

In the forward diffusion process (i.e., the noising process), we gradually add noise to the input image x:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \quad t \in \{1, \ldots, T\}, \tag{233}$$

where t is the discrete time step and $x_0 = x$. $\{\beta_t\}_{t=1}^T$ are the set of variance scheduling parameters, where $0 \leq \beta_t \leq 1$ and $\beta_{t-1} < \beta_t$. That is, we gradually add *more* noise to each of the intermediate images. ϵ is the standard Gaussian noise sampled from $\mathcal{N}(0, I)$. Denoting $\alpha_t = 1 - \beta_t$, we expand Eq. (233) recursively:

$$\begin{aligned} x_t =& \sqrt{\alpha_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \\ =& \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}x_{t-2} + \sqrt{\beta_{t-1}}\epsilon_{t-1}\right) + \sqrt{\beta_t}\epsilon_t \\ =& \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t, \end{aligned} \tag{234}$$

where we use the subscript t to differentiate Gaussian noises at different time steps, all from the same distribution. Expanding Eq. (234) further, we obtain

$$\begin{aligned} x_t =& \sqrt{\alpha_t\alpha_{t-1}}x_{t-2} + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t \\ =& \sqrt{\alpha_t\alpha_{t-1}}\left(\sqrt{\alpha_{t-2}}x_{t-3} + \sqrt{\beta_{t-2}}\epsilon_{t-2}\right) + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t \\ =& \sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}}x_{t-3} + \sqrt{\alpha_t\alpha_{t-1}\beta_{t-2}}\epsilon_{t-2} + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t \\ =& \cdots \\ =& \sqrt{\alpha_t \ldots \alpha_1}x_0 + \underbrace{\sqrt{\alpha_t \ldots \alpha_2\beta_1}\epsilon_1 + \sqrt{\alpha_t \ldots \alpha_3\beta_2}\epsilon_2 + \ldots + \sqrt{\alpha_t\beta_{t-1}}\epsilon_{t-1} + \sqrt{\beta_t}\epsilon_t}_{\text{sum of } t \text{ independent Gaussian noise samples}}. \end{aligned} \tag{235}$$

It is immediately clear that the term above the bracket is the sum of t independent noise samples from t Gaussian distributions with means zero and variances $\alpha_t \ldots \alpha_2\beta_1$, $\alpha_t \ldots \alpha_3\beta_2$, $\ldots$, $\alpha_t\beta_{t-1}$, and β_t, respectively. The additivity of independent Gaussian random variables[25] says that the sum of independent Gaussian random variables is still Gaussian with mean and variance equal to the sum of their respective means and variances. In other words, the term above the bracket is equivalent to and can be replaced by a Gaussian noise sample $\sqrt{\alpha_t \ldots \alpha_2\beta_1 + \alpha_t \ldots \alpha_3\beta_2 + \ldots + \alpha_t\beta_{t-1} + \beta_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, i.e.,

$$x_t = \sqrt{\alpha_t \ldots \alpha_1}x_0 + \sqrt{\alpha_t \ldots \alpha_2\beta_1 + \alpha_t \ldots \alpha_3\beta_2 + \ldots + \alpha_t\beta_{t-1} + \beta_t}\epsilon. \tag{236}$$

Note that

$$\begin{aligned}
&\alpha_t \ldots \alpha_2\beta_1 + \alpha_t \ldots \alpha_3\beta_2 + \ldots + \alpha_t\beta_{t-1} + \beta_t \\
=&\alpha_t \ldots \alpha_3(\alpha_2\beta_1 + \beta_2) + \alpha_t \ldots \alpha_4\beta_3 + \ldots + \alpha_t\beta_{t-1} + \beta_t \\
=&\alpha_t \ldots \alpha_3(\alpha_2(1-\alpha_1) + 1 - \alpha_2) + \alpha_t \ldots \alpha_4\beta_3 + \ldots + \alpha_t\beta_{t-1} + \beta_t \\
=&\alpha_t \ldots \alpha_3(1 - \alpha_1\alpha_2) + \alpha_t \ldots \alpha_4\beta_3 + \ldots + \alpha_t\beta_{t-1} + \beta_t \\
=&\alpha_t \ldots \alpha_4(\alpha_3(1 - \alpha_1\alpha_2) + \beta_3) + \alpha_t \ldots \alpha_5\beta_4 + \ldots + \alpha_t\beta_{t-1} + \beta_t \\
=&\alpha_t \ldots \alpha_4(1 - \alpha_1\alpha_2\alpha_3) + \alpha_t \ldots \alpha_5\beta_4 + \ldots + \alpha_t\beta_{t-1} + \beta_t \\
=&\cdots \\
=&\alpha_t(1 - \alpha_1\alpha_2 \ldots \alpha_{t-1}) + \beta_t \\
=&1 - \alpha_1\alpha_2 \ldots \alpha_t,
\end{aligned} \tag{237}$$

where we simply exploit the fact that $\alpha_t + \beta_t = 1$. Combining Eqs. (236) and (237), we have

$$\begin{aligned}
x_t =&\sqrt{\alpha_t \ldots \alpha_1}x_0 + \sqrt{\alpha_t \ldots \alpha_2\beta_1 + \alpha_t \ldots \alpha_3\beta_2 + \ldots + \alpha_t\beta_{t-1} + \beta_t}\epsilon \\
=&\sqrt{\alpha_1 \ldots \alpha_t}x_0 + \sqrt{1 - \alpha_1 \ldots \alpha_t}\epsilon \\
=&\sqrt{\bar{\alpha}_t} + \sqrt{1 - \bar{\alpha}_t}\epsilon,
\end{aligned} \tag{238}$$

where we define $\bar{\alpha}_t = \alpha_1 \ldots \alpha_t$.

10.2 On the Reverse Diffusion

We now justify the parameterization of $q_\theta\left(x_{t-1}|x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \sigma_t^2 I\right)$, in which

$$\mu_\theta\left(x_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t\right)\right), \tag{239}$$

according to the forward diffusion step $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$.

[25] https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

Proof. We first rewrite $q\left(x_{t-1}|x_t\right)$ using the Bayes' rule:

$$\begin{aligned}
q\left(x_{t-1}|x_t\right) &= q\left(x_{t-1}|x_t, x_0\right) \\
&= \frac{q\left(x_t, x_{t-1}|x_0\right)}{q\left(x_t|x_0\right)} \\
&= \frac{q\left(x_t|x_{t-1}, x_0\right) q\left(x_{t-1}|x_0\right)}{q\left(x_t|x_0\right)} \\
&= \frac{q\left(x_t|x_{t-1}\right) q\left(x_{t-1}|x_0\right)}{q\left(x_t|x_0\right)} \\
&= \frac{\mathcal{N}\left(x_t; \sqrt{\alpha_t} x_{t-1}, (1-\alpha_t) I\right) \mathcal{N}\left(x_{t-1}; \sqrt{\bar{\alpha}_{t-1}} x_0, \left(1-\bar{\alpha}_{t-1}\right) I\right)}{\mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, \left(1-\bar{\alpha}_t\right) I\right)} \\
&= \mathcal{N}\left(x_{t-1}; \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t} x_0, \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} I\right). \qquad (240)
\end{aligned}$$

In the first equation, we write out the dependence of x_{t-1} on x_0 because x_{t-1} is conditioned on x_t, which is in turn dependent on x_0 through the forward diffusion step $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$. From the third to the fourth equation, we drop the condition of x_0 in $q(x_t|x_{t-1})$ due to the Markovity of the forward diffusion process: x_0 and x_t are conditional independent given x_{t-1}. The last equation is the result of the *completing the square* trick. Specifically, we focus on the terms in the exponent:

$$E_{\text{term}} = \frac{\|x_t - \sqrt{\alpha_t} x_{t-1}\|_2^2}{1-\alpha_t} + \frac{\|x_{t-1} - \sqrt{\bar{\alpha}_{t-1}} x_0\|_2^2}{1-\bar{\alpha}_{t-1}} - \frac{\|x_t - \sqrt{\bar{\alpha}_t} x_0\|_2^2}{1-\bar{\alpha}_t}, \qquad (241)$$

which is quadratic in x_{t-1}, meaning that x_{t-1} has to be Gaussian distributed. We gather the quadratic terms of x_{t-1}:

$$\begin{aligned}
Q_{\text{term}} &= \left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right) \|x_{t-1}\|_2^2 \\
&= \frac{\alpha_t(1-\bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} \|x_{t-1}\|_2^2 \\
&= \frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} \|x_{t-1}\|_2^2, \qquad (242)
\end{aligned}$$

where we make use of $\bar{\alpha}_t = \bar{\alpha}_{t-1}\alpha_t = \alpha_1 \ldots \alpha_t$ in the second equation. We next gather the linear terms of x_{t-1}:

$$\begin{aligned}
L_{\text{term}} &= -2\left(\frac{\sqrt{\alpha_t}}{1-\alpha_t} x_t^T + \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_{t-1}} x_0^T\right) x_{t-1} \\
&= -2\left(\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t^T + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t) x_0^T}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right) x_{t-1}. \qquad (243)
\end{aligned}$$

Now it is ready to complete the square:

$$E_{\text{term}} = \frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})} \left\| x_{t-1} - \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1}) x_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t) x_0}{1-\bar{\alpha}_t} \right\|_2^2 + \text{const}, \qquad (244)$$

where "const" denotes all remaining terms irrelevant to x_{t-1}. From Eq. (244), it is straightforward to derive the mean $\mu_\theta\left(x_t, t\right)$ and the variance σ_t^2 of $q_\theta\left(x_{t-1}|x_t\right)$:

$$\mu_\theta(x_t, t) = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0, \tag{245}$$

and

$$\sigma_t^2 = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}, \tag{246}$$

as desired. Plugging the forward diffusion step $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ into Eq. (245), we obtain

$$\begin{aligned}
\mu_\theta(x_t, t) =& \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon\right) \\
=& \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{(1-\alpha_t)}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon\right) \\
=& \frac{\alpha_t\,(1-\bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{(1-\bar{\alpha}_t)\alpha_t}}\epsilon \\
=& \frac{1-\bar{\alpha}_t}{(1-\bar{\alpha}_t)\sqrt{\alpha_t}}x_t - \frac{1-\alpha_t}{\sqrt{(1-\bar{\alpha}_t)\alpha_t}}\epsilon \\
=& \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon\right),
\end{aligned} \tag{247}$$

which completes the justification of the parameterization in Eq. (239).

□