Machine Learning

Summary by Ari Feiglin (ari.feiglin@gmail.com)

Contents

1 Bayes Decision Theory

1.1 Formalizing the Theory

Let us imagine the following scenario: you, a computer science student, are preparing to leave your house for the first time in a while. Should you or should you not take your umbrella? If you take your umbrella and it doesn't rain then you've inconvenienced yourself, but if you don't and it rains then you'll end up getting wet. You look outside, see that clouds are grey, and decide to take your umbrella.

Here you are trying to decide between some possible actions, each with their own cost. These actions are based on the state of the world, of which you have some set of observations.

The problem can be formalized as follows, you have

(1) a set of world states: $\{\omega_i\}_{i \in I}$, which are disjoint and exhaustive: $\Omega = \bigcup_{i \in I} \omega_i$ (where Ω is the entire space).

(2) a set of observations: $S_n = \{x_1, \ldots, x_n\}$.

(3) a probabilistic model: conditionals $\mathbb{P}(S_n \mid \omega)$ and priors $\mathbb{P}(\omega)$.

(4) a set of possible actions $A = \{\alpha_1, \ldots, \alpha_k\}$.

(5) a set of cost functions $\Lambda = \{\lambda(\alpha_k \mid \omega_j)\}_{j,k}$.

So in our example above, we have two world states: raining and not raining, our observation is that the clouds are grey, we have some prior belief about the probabilities of each world state as well as the probability of observing our observation given a world state, our actions are to take and to not take an umbrella, each has an associated cost.

So suppose we've made an observation x, we'd like to compute the new probability of a world state ω given this observation. This can be done via Bayes's law:

$$\mathbb{P}(\omega \mid x) = \frac{\mathbb{P}(x \mid \omega)}{\mathbb{P}(x)} \cdot \mathbb{P}(\omega)$$

$\mathbb{P}(\omega \mid x)$ is called the *posterior*. But $\mathbb{P}(x)$ is not known, so how do we compute it? Using total probability:

$$\mathbb{P}(x) = \sum_{i \in I} \mathbb{P}(x \mid \omega_i)\, \mathbb{P}(\omega_i)$$

As we make more and more observations, we can employ Baye's law over and over to refine the posterior.

1.2 Minimizing Risk

Our goal in Bayes decision theory is to define a strategy $\alpha(S_n)$ which determines which action α to take so that it minimizes our expected costs, given observations S_n.

Definition 1.2.1

Given an action α, we define its **conditional risk** given an observation x to be

$$R(\alpha \mid x) = \sum_{i \in I} \lambda(\alpha \mid \omega_i)\, \mathbb{P}(\omega_i \mid x)$$

Where the posterior $\mathbb{P}(\omega_i \mid x)$ is computed as above using Bayes's law. If we define the random variable $\lambda(\alpha)$ to be $\lambda(\alpha \mid \omega)$ under the state ω, then this is just $\mathbb{E}[\lambda(\alpha) \mid x]$.

Example 1.2.2

Suppose our actions $\alpha_1, \ldots, \alpha_k$ are to guess the world state $\omega_1, \ldots, \omega_k$. The cost we pay is 1 if we are wrong and 0 if we are correct, meaning $\lambda(\alpha_k \mid \omega_j) = 1 - \delta_{kj}$ where $\delta_{kj} = 1$ when $k = j$ and 0 otherwise.

Then the conditional risk is

$$R(\alpha_k \mid x) = \sum_{j=1}^{k} \lambda(\alpha_k \mid \omega_j)\,\mathbb{P}(\omega_j \mid x) = \sum_{j \neq k} \mathbb{P}(\omega_j \mid x) = 1 - \mathbb{P}(\omega_k \mid x)$$

So we minimize the conditional risk when we take α_k such that the posterior $\mathbb{P}(\omega_k \mid x)$ is maximal.

Example 1.2.3

Suppose we have two world states ω_1, ω_2 and two actions α_1, α_2. Then our costs form a 2×2 matrix: $\lambda_{kj} = \lambda(\alpha_k \mid \omega_j)$. And by definition

$$R(\alpha_i \mid x) = \lambda_{i1}\,\mathbb{P}(\omega_1 \mid x) + \lambda_{i2}\,\mathbb{P}(\omega_2 \mid x)$$

We choose α_1 if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$, which is equivalent to

$$(\lambda_{12} - \lambda_{22})\,\mathbb{P}(\omega_2 \mid x) < (\lambda_{21} - \lambda_{11})\,\mathbb{P}(\omega_1 \mid x)$$

which is equivalent to

$$\iff \frac{\mathbb{P}(\omega_1 \mid x)}{\mathbb{P}(\omega_2 \mid x)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \iff \frac{\mathbb{P}(x \mid \omega_1)}{\mathbb{P}(x \mid \omega_2)} > \frac{\mathbb{P}(\omega_2)}{\mathbb{P}(\omega_1)} \cdot \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$$

The left-hand side is called the **likelihood ratio** and the right-hand side is called the **decision boundary**.

If we have many observations $x_1, \ldots, x_n$ then this becomes

$$\overset{\text{Likelihood ratio}}{\frac{\mathbb{P}(x_1, \ldots, x_n \mid \omega_1)}{\mathbb{P}(x_1, \ldots, x_n \mid \omega_2)}} > \overset{\text{Decision boundary}}{\frac{\mathbb{P}(\omega_2)}{\mathbb{P}(\omega_1)} \cdot \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}}} = \Theta$$

If the observations $x_1, \ldots, x_n$ are independent then this becomes

$$\frac{\prod_i \mathbb{P}(x_i \mid \omega_1)}{\prod_i \mathbb{P}(x_i \mid \omega_2)} > \Theta$$

taking the log of both sides gives

$$\sum_i \log \frac{\mathbb{P}(x_i \mid \omega_1)}{\mathbb{P}(x_i \mid \omega_2)} > \log \Theta = \Theta'$$

the left-hand side is called the **log-likelihood ratio**.

1.3 Parameter Estimation

Suppose we know the distribution of some random variable X up to some parameter θ, i.e. we know the function $\mathbb{P}(X \mid \theta)$. For example, X is the number of heads in n coin tosses where the coin has a bias of θ, then $X \mid \theta \sim \mathrm{Bin}(n, \theta)$.

We use Bayes decision theory to estimate θ. Suppose we have a prior $\mathbb{P}(\theta)$, we use this to estimate θ. Our actions will be choosing some prediction of θ, $\hat{\theta}$. Define a cost function $\lambda(\hat{\theta}, \theta)$. Then given a sequence of observations $S_n = \{x_1, \ldots, x_n\}$, we want to minimize the expected cost

$$\mathbb{E}[\lambda(\hat{\theta}) \mid S_n] = \int \lambda(\hat{\theta}, \theta)\,\mathbb{P}(\theta \mid S_n)\,d\theta$$

We then define the *Bayes estimator* to be the prediction θ^* which minimizes this:

$$\theta^* = \mathrm{argmin}_{\hat{\theta}}\,\mathbb{E}[\lambda(\hat{\theta}) \mid S_n]$$

To find θ^* we differentiate $\mathbb{E}[\lambda(\hat{\theta}) \mid S_n]$ and compare it with zero. Hand-waving away all the technical details because this is computer science, we can swap the order of differentiation and integration and so

$$\frac{d}{d\hat{\theta}} \mathbb{E}[\lambda(\hat{\theta}) \mid S_n] = \frac{d}{d\hat{\theta}} \int \lambda(\hat{\theta}, \theta)\, \mathbb{P}(\theta \mid S_n)\, d\theta = \int \frac{d}{d\hat{\theta}} \lambda(\hat{\theta}, \theta)\, \mathbb{P}(\theta \mid S_n)\, d\theta$$

Example 1.3.1 (Square Loss)

Suppose we use a **square loss** cost function: $\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$. Then the Bayes estimator is

$$\mathbb{E}[\lambda(\hat{\theta})] = \int \lambda(\hat{\theta}, \theta)\, \mathbb{P}(\theta \mid S_n)\, d\theta = \int (\theta - \hat{\theta})^2\, \mathbb{P}(\theta \mid S_n)\, d\theta$$

Differentiating gives that we want

$$\int \theta\, \mathbb{P}(\theta \mid S_n)\, d\theta = \hat{\theta} \int \mathbb{P}(\theta \mid S_n)\, d\theta = \hat{\theta}$$

(since $\int \mathbb{P}(\theta \mid S_n) = 1$.) So we get that our estimator is

$$\hat{\theta}_{SE} = \int \theta\, \mathbb{P}(\theta \mid S_n)\, d\theta = \mathbb{E}[\theta \mid S_n]$$

Example 1.3.2 (Maximum Aposteriori)

Suppose we use a **zero-one** cost function: $\lambda(\hat{\theta}, \theta) = 1 - \delta_{\hat{\theta}\theta}$. We have already showed that to minimize the cost, we must maximize $\operatorname{argmax} \mathbb{P}(\theta \mid S_n)$. By Bayes, this is just $\operatorname{argmax} \mathbb{P}(S_n \mid \theta) \frac{\mathbb{P}(\theta)}{\mathbb{P}(S_n)} = \operatorname{argmax} \mathbb{P}(S_n \mid \theta)\, \mathbb{P}(\theta)$. Since the observations in S_n are independent, we see that this is then just equal to $\operatorname{argmax} \prod_i \mathbb{P}(x_i \mid \theta)\, \mathbb{P}(\theta)$. Finally we take the log, which is monotonic, to get that the estimator

$$\hat{\theta}_{MAP} = \operatorname{argmax} \sum_i \log \mathbb{P}(x_i \mid \theta) + \log \mathbb{P}(\theta)$$

as n grows, the last term becomes negligible and this just becomes the maximum log likelihood function.

Exercise 1.3.3

Compute the MAP estimator for $X \sim \mathrm{Exp}(\theta)$, i.e. $f_X(x) = \theta \exp(-\theta x)$, with the prior $\mathbb{P}(\theta) = \exp(-\theta)$. We want to compute

$$\sum_i \log \mathbb{P}(x_i \mid \theta) + \log \mathbb{P}(\theta) = \sum_i \log\bigl(\theta \exp(-\theta x_i)\bigr) + \log \exp(-\theta) = n \log \theta - \theta \sum_i x_i - \theta$$

Differentiating with respect to θ gives

$$\frac{n}{\theta} - \sum_i x_i - 1 = 0 \implies \hat{\theta}_{MAP} = \frac{n}{\sum_i x_i + 1}$$

Here the prior behaves similarly to a sample, it can be viewed as a "pseudo-sample".

Example 1.3.4 (Maximum Likelihood)

We know that

$$\hat{\theta}_{MAP} = \operatorname{argmax} \sum_i \log \mathbb{P}(x_i \mid \theta) + \log \mathbb{P}(\theta)$$

A common approach is to approximate this by dropping the prior, giving

$$\hat{\theta}_{ML} = \operatorname{argmax} \sum_i \log \mathbb{P}(x_i \mid \theta)$$

Asymptotically this is efficient.

2 PAC Learning

2.1 Definitions

We now define *Probably Approximately Correct (PAC)* Learning.

Definition 2.1.1

Let $\mathcal{X}$ be the **input space**, the set of all possible examples or instances. The set of all **labels** or **target values** is $\mathcal{Y}$. For now, we restrict our view to be binary: $\mathcal{Y} = \{0, 1\}$. A **concept** is a map $c\colon \mathcal{X} \longrightarrow \mathcal{Y}$, or equivalently a subset of $\mathcal{X}$ (the set $\{x \in \mathcal{X} \mid c(x) = 1\}$). A **concept class** is a class of concepts which we would like to learn (approximate) and is denoted $\mathcal{C}$.

The idea of PAC learning is as follows: the learner considers a fixed set of concepts $\mathcal{H}$ called the *hypothesis set*, which may or may not coincide with $\mathcal{C}$. The learner then receives a sequence of samples $S = (x_1, \dots, x_n)$ which are independent and distribute according to some distribution $\mathcal{D}$. The learner also receives labels $(c(x_1), \dots, c(x_n))$ according to some concept $c \in \mathcal{C}$ which it is tasked with learning. Using this information the learner attempts to choose a hypothesis $h_S \in \mathcal{H}$ which minimizes the *generalization error* (or *risk*):

Definition 2.1.2

Given a hypothesis $h \in \mathcal{H}$, target concept $c \in \mathcal{C}$, and an underlying distribution $\mathcal{D}$, the **generalized error** (or **risk**) is:

$$R(h) = \mathbb{P}(h(x) \neq c(x) \mid x \sim \mathcal{D}) = \mathbb{E}[\mathbf{1}\{h(x) \neq c(x)\} \mid x \sim \mathcal{D}]$$

i.e. it is the probability that $h(x)$ differs from $c(x)$ when x is chosen randomly with a distribution of $\mathcal{D}$.

But the generalized error cannot be known to the learner, as it knows not the target concept nor the underlying distribution. So instead the learner minimizes the *empirical error* (or *risk*):

Definition 2.1.3

Given a hypothesis $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, and a sample $S = (x_1, \dots, x_n)$, define the **empirical error** (or **risk**) to be:

$$\widehat{R}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq c(x_i)\}$$

Notice that

$$\begin{aligned}\mathbb{E}[\widehat{R}_S(h) \mid S \sim \mathcal{D}^n] &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{1}\{h(x_i) \neq c(x_i)\} \mid S \sim \mathcal{D}^n] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbf{1}\{h(x) \neq c(x)\} \mid x \sim \mathcal{D}] = \frac{1}{n} \sum_{i=1}^{n} R(h) = R(h)\end{aligned}$$

We now formally define what PAC learning is. Let n be a number such that the size of every $x \in \mathcal{X}$ can be represented in $O(n)$ space, for $c \in \mathcal{C}$ let $\mathit{size}\, c$ be the maximal computational cost of c. We focus on algorithms $\mathcal{A}$ which take as input a sample S and return a hypothesis h_S.

Definition 2.1.4

A concept class $\mathcal{C}$ is **PAC-learnable** if there exists an algorithm $\mathcal{A}$ and a polynomial function $\mathit{poly}(\bullet, \bullet, \bullet, \bullet)$ such that for all $\varepsilon, \delta > 0$, distribution $\mathcal{D}$ on $\mathcal{X}$ and target concept $c \in \mathcal{C}$, for every sample size $n \geq \mathit{poly}(1/\varepsilon, 1/\delta, n, \mathit{size}\, c)$,

$$\mathbb{P}(R(h_S) \leq \varepsilon \mid S \sim \mathcal{D}^n) \geq 1 - \delta$$

If $\mathcal{A}$ runs in $\mathit{poly}(1/\varepsilon, 1/\delta, n, \mathit{size}\, c)$ time then $\mathcal{C}$ is **efficiently PAC-learnable**. An algorithm $\mathcal{A}$, if one exists, is called a **PAC-learning algorithm** for $\mathcal{C}$.

The intuition is as follows: a concept class $\mathcal{C}$ is PAC-learnable if there exists an algorithm $\mathcal{A}$ where given a sample size at least polynomial in $1/\varepsilon$ and $1/\delta$, it returns a hypothesis with an error bound by ε at least $1-\delta$ of the time. Note that if the running time is polynomial in $1/\varepsilon$ and $1/\delta$, then assuming the total input is read by the algorithm, the input must too be polynomial in $1/\varepsilon$ and $1/\delta$.

2.2 The Finite Case

Given a concept $c \in \mathcal{C}$ and a sample $S = (x_1, \dots, x_n)$, call an hypothesis $h \in \mathcal{H}$ *consistent* if $h(x_i) = c(x_i)$ for all $1 \leq i \leq n$. Equivalently, $\widehat{R}_S(h) = 0$. We will assume that our hypotheses are consistent, and so we can always assume that our target concept is in $\mathcal{H}$.

Theorem 2.2.1

Let $\mathcal{H}$ be a finite set of hypotheses, and $\mathcal{A}$ an algorithm such that for any target concept $c \in \mathcal{H}$, $\mathcal{A}$ returns a consistent hypothesis h_S for any input sample S (where $S \sim \mathcal{D}^n$). Then for every every $\varepsilon, \delta > 0$, the inequality $\mathbb{P}(R(h_S) \leq \varepsilon \mid S \sim \mathcal{D}^n) \geq 1 - \delta$ holds if

$$n \geq \frac{1}{\varepsilon} \cdot \log \frac{|\mathcal{H}|}{\delta}$$

Proof: let $\varepsilon > 0$, and define $\mathcal{H}_\varepsilon = \{h \in \mathcal{H} \mid R(h) > \varepsilon\}$. The probability that $h \in \mathcal{H}_\varepsilon$ is consistent is

$$\begin{aligned}\mathbb{P}(\widehat{R}_S(h) = 0) &= \mathbb{P}(h(x_1) = c(x_1), \dots, h(x_n) = c(x_n) \mid x_i \sim \mathcal{D}) \\ &= \prod_{i=1}^{n} \mathbb{P}(h(x_i) = c(x_i) \mid x_i \sim \mathcal{D}) = \mathbb{P}(h(x) = c(x) \mid x \sim \mathcal{D})^n = \big(1 - \mathbb{P}(h(x) \neq c(x) \mid x \sim \mathcal{D})\big)^n\end{aligned}$$

this is since the samples are independent and distributively equal. Now recall that by definition, we have that $\mathbb{P}(h(x) \neq c(x) \mid x \sim \mathcal{D}) = R(h)$ which is greater than ε by definition, so

$$\mathbb{P}(\widehat{R}_S(h) = 0) \leq (1 - \varepsilon)^n$$

Thus the probability that a "bad" hypothesis tricks us by being consistent is bound by $(1-\varepsilon)^n$. Now, we want to show that the probability that there exists a bad consistent hypothesis is bound by δ. This means that with probability at least $1 - \delta$, there exist no bad consistent hypotheses and so h_S is necessarily a "good" hypothesis with $R(h_S) \leq \varepsilon$. The probability of there existing a consistent bad hypothesis is

$$\mathbb{P}(\exists h \in \mathcal{H}_\varepsilon .\ \widehat{R}_S(h) = 0) \leq \sum_{h \in \mathcal{H}_\varepsilon} \mathbb{P}(\widehat{R}_S(h) = 0) \leq \sum_{h \in \mathcal{H}_\varepsilon} (1-\varepsilon)^n \leq |\mathcal{H}|(1-\varepsilon)^n$$

We also know that $(1-\varepsilon)^n \leq e^{-n\varepsilon}$, so this probability is bound by $|\mathcal{H}|e^{-n\varepsilon}$. If $n \geq \frac{1}{\varepsilon} \cdot \log \frac{|\mathcal{H}|}{\delta}$ then this is bound by δ, as required. ■

Since $\frac{1}{\varepsilon} \cdot \log \frac{|\mathcal{H}|}{\delta}$ is bound by some polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta}$, this means that when $\mathcal{H}$ is finite a consistent algorithm $\mathcal{A}$ is PAC-learning. But what if our algorithm isn't consistent, i.e. the target concept is not in $\mathcal{H}$?

Theorem 2.2.2

For every $\varepsilon, \delta > 0$ if $n > \frac{1}{\varepsilon^2} \log \frac{2|\mathcal{H}|}{\delta}$ then

$$\mathbb{P}\left(R(h_S) - \min_{h \in \mathcal{H}} R(h) \leq \varepsilon\right) \geq 1 - \delta$$

Proof: similar to before, but using Hoeffding's inequality $\mathbb{P}(S - \mathbb{E}[S] \geq \varepsilon) \leq \exp(-2\varepsilon^2/n)$. ■

This is a generalization of the previous theorem, since if c is taken from $\mathcal{H}$ then $\min_{h \in \mathcal{H}} R(h) = 0$. Call a concept class $\mathcal{C}$ which satisfies this probability bound *Agonistically PAC-learnable*:

Definition 2.2.3

A concept class $\mathcal{C}$ is **agonistically PAC-learnable** if there exists an **agnostic PAC-learner** $\mathcal{A}$ such that for every $\varepsilon, \delta > 0$ distribution $\mathcal{D}$ and $n \geq \mathit{poly}(1/\varepsilon, 1/\delta, n, \mathit{size}\, c)$ for some *poly*,

$$\mathbb{P}\left(R(h_S) - \min_{h\in\mathcal{H}} R(h) \leq \varepsilon\right) \geq 1 - \delta$$

So assuming the hypothesis class is finite, every concept class is PAC-learnable in the agnostic sense.

2.3 The Infinite Case

Suppose now that $\mathcal{H}$ is infinite, and we don't necessarily have that the target concept is in $\mathcal{H}$. Define

$$h_S := \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{R}_S(h), \qquad h_{\mathcal{D}} := \operatorname*{argmin}_{h\in\mathcal{H}} R_{\mathcal{D}}(h)$$

Then we have

Lemma 2.3.1

$$R_{\mathcal{D}}(h_S) - R_{\mathcal{D}}(h_{\mathcal{D}}) \leq 2 \sup_{h\in\mathcal{H}} |\widehat{R}_S(h) - R_{\mathcal{D}}(h)|$$

Proof: let $\varepsilon = \sup_{h\in\mathcal{H}} |\widehat{R}_S(h) - R_{\mathcal{D}}(h)|$, then

$$\begin{aligned} R_{\mathcal{D}}(h_S) &\leq \widehat{R}_S(h_S) + \varepsilon && \text{since } h_S \in \mathcal{H} \\ &\leq \widehat{R}_S(h_{\mathcal{D}}) + \varepsilon && \text{since } h_S \text{ minimizes } \widehat{R}_S \\ &\leq R_{\mathcal{D}}(h_{\mathcal{D}}) + 2\varepsilon && \text{since } h_{\mathcal{D}} \in \mathcal{H} \end{aligned}$$

■

Definition 2.3.2

A hypothesis class $\mathcal{H}$ **shatters** a finite set $C = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ if for any labels $y_1, \ldots, y_n \in \{0, 1\}$ there exists an hypothesis $h \in \mathcal{H}$ such that $h(x_i) = y_i$ for all i.

In other words, $\mathcal{H}$ shatters C if for every function $f \colon C \longrightarrow \mathcal{Y}$, there exists an hypothesis h such that $h\big|_C = f$. So if we define $\mathcal{H}_C = \left\{h\big|_C \mid h \in \mathcal{H}\right\}$, then $\mathcal{H}$ shatters C if and only if $|\mathcal{H}_C| = 2^{|C|}$.

Definition 2.3.3

The **VC dimension** of an hypothesis class $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the size of the largest set $C \subseteq \mathcal{X}$ shattered by $\mathcal{H}$.

Theorem 2.3.4

Let $\mathcal{H}$ be an hypothesis class with VC dimension $d < \infty$. Then there exists a constant $c > 0$ such that for every $\varepsilon, \delta \in (0, 1)$ there is a polynomial such that for every $n \geq \mathit{poly}(1/\varepsilon, 1/\delta)$ for which

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} |\widehat{R}_S(h) - R_{\mathcal{D}}(h)| < c\sqrt{\frac{d}{n}} \;\middle|\; S \sim \mathcal{D}^n\right) \geq 1 - \delta$$

The proof of this is beyond the scope of the class.

3 Linear Regression

3.1 Linear Regression

Suppose our input space $\mathcal{X}$ is a subset of $\mathbb{R}^d$ for some d, and $\mathcal{Y}$ is $\mathbb{R}$. Given a sample $\{(\vec{x}_i, y_i)\}_{1\le i\le n}$, we would like to find the best linear approximation $h\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ which approximates the relation between the inputs and their labels. Thus the hypothesis class is then the set of all linear functions

$$\mathcal{H} = \left\{ f(\vec{x}) = \vec{x}^\top \vec{w} + b \;\middle|\; \vec{w} \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

For a hypothesis f, let us denote $\hat{y}_i = f(x_i) = \vec{x}_i^\top \vec{w} + b$, notice that

$$\begin{pmatrix} \text{———} & \vec{x}_1 & \text{———} & 1 \\ & \vdots & & \vdots \\ \text{———} & \vec{x}_n & \text{———} & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_d \\ b \end{pmatrix} = \begin{pmatrix} \vec{x}_1^\top \vec{w} + b \\ \vdots \\ \vec{x}_n^\top \vec{w} + b \end{pmatrix}$$

To simplify notation, we can just assume that the final coordinate of each $\vec{x}_i$ is 1 and so we can view each hypothesis f as its vector $\vec{w}$ (where the last coefficient is b). Thus

$$X\vec{w} = \hat{y}$$

Where X is the matrix whose rows are $\vec{x}_i$.

We want to minimize the difference between $\hat{y}$ and $\vec{y}$. We can do this by taking the mean square error (MSE): $\frac{1}{n}\|\hat{y} - \vec{y}\|^2 = \frac{1}{n}\|X\vec{w} - \vec{y}\|^2$. So we define

$$\mathrm{MSE}(\vec{w}) = \frac{1}{n}\|X\vec{w} - \vec{y}\|^2 = \frac{1}{n}(X\vec{w} - \vec{y})^\top (X\vec{w} - \vec{y})$$

We take the gradient and compare with zero:

$$\nabla \mathrm{MSE}(\vec{w}) = \frac{2}{n} X^\top (X\vec{w} - \vec{y})$$

comparing with zero gives

$$\nabla \mathrm{MSE}(\vec{w}) = 0 \iff X^\top X\vec{w} - X^\top \vec{y} = 0 \iff \vec{w}_{\min} = (X^\top X)^{-1} X^\top \vec{y}$$

This must be a minimum since MSE is quadratic in $\vec{w}$.

So now if we have new data X_{new}, our estimator for the new values will be

$$\hat{y}_{\text{new}} = X_{\text{new}} \vec{w}_{\min} = X_{\text{new}} (X^\top X)^{-1} X^\top \vec{y}$$

Furthermore, notice that the error $\vec{y} - \hat{y}$ is orthogonal to $\hat{y}$:

$$\begin{aligned}(\vec{y} - \hat{y})^\top \hat{y} = (\vec{y} - X\vec{w})^\top \hat{y} &= \vec{y}^\top \hat{y} - \vec{w}^\top X^\top X\vec{w} \\ &= \vec{y}^\top X\vec{w} - \vec{w}^\top X^\top X (X^\top X)^{-1} X^\top \vec{y} \\ &= \vec{y}^\top X\vec{w} - \vec{w}^\top X^\top \vec{y} \\ &= \vec{y}^\top X\vec{w} - (X\vec{w})^\top \vec{y} = 0\end{aligned}$$

3.2 Polynomial Fitting

Suppose we have n points $\{(x_i, y_i)\}_{1\le i\le n}$ which we would like to approximate using a k-degree polynomial $\hat{y} = p(x) = w_0 x^0 + \cdots + w_k x^k$. We can do so using linear regression: define

$$X = \begin{pmatrix} x_1^0 & x_1^1 & \cdots & x_1^k \\ \vdots & \vdots & & \vdots \\ x_n^0 & x_n^1 & \cdots & x_n^k \end{pmatrix}$$

Then doing linear regression gets us the best vector $\vec{w}$ such that

$$X\vec{w} = \begin{pmatrix} \sum_{i=0}^{k} w_i x_1^i \\ \vdots \\ \sum_{i=0}^{k} w_i x_n^i \end{pmatrix} = \begin{pmatrix} p(x_1) \\ \vdots \\ p(x_n) \end{pmatrix}$$

is closest to $\vec{y}$, as required.

3.3 Ridge Regression

What if $X^\top X$ is not invertible? Which can happen if we have too few samples, i.e. $n < d$. Notice that $X^\top X$ is positive semi-definite: $v^\top X^\top X v = (Xv)^\top (Xv) \geq 0$, and thus all its eigenvalues are nonnegative. Then for any $\lambda > 0$, $(X^\top X + \lambda I)v = \mu v$ if and only if $X^\top X v = (\mu - \lambda)v$ which must mean that $\mu - \lambda \geq 0$ and in particular $\mu > 0$. So all of $X^\top X + \lambda I$'s eigenvalues are positive, and in particular non-zero, meaning it is invertible.

So using this knowledge, let us define the *ridge estimator* to be

$$\widehat{w}_{\text{ridge}} = \underset{\vec{w}}{\operatorname{argmin}} \|X\vec{w} - \vec{y}\|^2 + \lambda \|\vec{w}\|^2$$

Taking the gradient and equating to zero gives that

$$\widehat{w}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$$

which, as explained above, exists no matter what, for any $\lambda > 0$.

4 Binary Classification

Suppose we'd like to learn the class of halfspaces: $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{-1, 1\}$ and the hypothesis class is

$$\mathcal{HS}_d = \left\{\mathbf{x} \mapsto \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b) \;\middle|\; \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\right\}$$

The idea here is that we split our space by the hyperplane $\langle \mathbf{w}, \mathbf{x} \rangle = -b$, and we classify elements as being either above or below the hyperplane depending on the sign of $\langle \mathbf{w}, \mathbf{x} \rangle + b$. Let us assume the realizable case, then we'd like to learn this hypothesis class.

Notice that

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = \langle (b, \mathbf{w}), (1, \mathbf{x}) \rangle$$

so we can assume that $b = 0$ and the input $\mathbf{x}$'s first index is 1.

4.1 The Perceptron Algorithm

```
function Perceptron($(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$)
    $\mathbf{w}^{(1)} \leftarrow (0, \dots, 0)$
    for ($t = 1, 2, 3, \dots$)
        if (There exists an $i$ such that $y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0$)
            $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
        else
            return $\mathbf{w}^{(t)}$
        end if
    end for
end function
```

Notice that the check $y_i \langle \mathbf{w}^{(t)}, \mathbf{x} \rangle \leq 0$ is the same as $\operatorname{sign}\langle \mathbf{w}^{(t)}, \mathbf{x} \rangle \neq y_i$. So this algorithm checks if the current $\mathbf{w}^{(t)}$ mislabels any input, and if so it alters $\mathbf{w}^{(t)}$ to fix this. Notice that $w^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$ will increase $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle$:

$$y_i \left\langle \mathbf{w}^{(t+1)}, \mathbf{x}_i \right\rangle = y_i \left\langle \mathbf{w}^{(t)} + y_i \mathbf{x}_i, \mathbf{x}_i \right\rangle = y_i \left\langle \mathbf{w}^{(t)}, \mathbf{x}_i \right\rangle + \|\mathbf{x}_i\|^2$$

so this will make the current hypothesis hopefully closer to objective concept.

Indeed this algorithm does terminate, but the proof of this is not relevant to this course.

4.2 Gradient Descent

Suppose we want to minimize some loss function $\ell(\mathbf{w})$. Since $\nabla \ell$ is the direction in which the loss function increases the most, if we go in the opposite direction we hope that this decreases the function. So if we start with an initial guess for the minimum point of the loss function $\mathbf{w}^{(1)}$, then at each iteration we just compute

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \nabla \ell(\mathbf{w}^{(t)})$$

where $\eta > 0$ is some constant. This iterative procedure is called *gradient descent*. Assuming ℓ is convex and $\nabla \ell$ is Lipschitz-continuous, and $0 < \eta < 1/L$ where L is the Lipschitz constant for $\nabla \ell$, then gradient descent is ensured to converge.

4.3 Stochastic Gradient Descent

When we have a collection of samples $\{(x_i, y_i)\}_{i=1}^n$, then we want to minimize $Err(w) = \sum_{i=1}^n \ell(w; x_i, y_i)$ (i.e. ℓ is parameterized, and we want to minimize the mean over the samples). Then our iteration becomes

$$w^{(t+1)} \leftarrow w^{(t)} - \eta \nabla Err(w^{(t)})$$

But computing the gradient of n parameterizations is costly, so we can approximate this by defining

$$\textit{Batch-Err}(w) = \sum_{i=1}^{b} \ell(w; x_i, y_i)$$

where $b \ll n$ and the parameters x_i, y_i are chosen randomly from the set of samples (the samples are chosen uniformly at each iteration). When $b = 1$ this is called *pure stochastic gradient descent*.

4.4 Stochastic Gradient Descent Learning

Let us return to the problem of learning halfspaces. Recall that our samples are of the form $\{(\mathbf{x}_i, y_i)\}$ where $y_i = \pm 1$ depending on whether or not $\mathbf{x}_i$ is in or out of the halfspace. The perceptron method works for the realizable case, but in the case that there is some noise in the sampling, it will not work. So let us use SGD to learn it, to do so we take a good loss function, for example the *hinge loss function*:

$$\ell_{hinge}(w; x, y) = \begin{cases} -yw^\top x & yw^\top x < 0 \\ 0 & \text{else} \end{cases}$$

Then the total loss function is

$$Err(w) = \sum_{i=1}^{n} \ell_{hinge}(w; \mathbf{x}_i, y_i)$$

We compute:

$$\nabla \ell_{hinge}(w; x, y) = \begin{cases} -yx & yw^\top x < 0 \\ 0 & \text{else} \end{cases}$$

So using pure SGD we get the following algorithm for learning halfspaces:

```
function SGD-LEARNER($(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), \eta$)
    $\mathbf{w}^{(1)} \leftarrow (0, \ldots, 0)$
    while (There exists an $i$ such that $y_i(\mathbf{w}^{(t)})^\top x_i < 0$)
        $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \eta y_i \mathbf{x}_i$
    end while
end function
```

4.5 Logistic Regression

The issue with the previous methods is that they check $\operatorname{sign}(\mathbf{w}^\top \mathbf{x}_i)$, which is binary and therefore very rigid. Instead, we can define the hypothesis class

$$\mathcal{H}_{\text{sig}} = \left\{ \mathbf{x} \mapsto \sigma(\langle \mathbf{w}, \mathbf{x} \rangle) \;\middle|\; \mathbf{w} \in \mathbb{R}^d \right\}$$

where σ is the *sigmoid function*:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

The shape of this graph looks a lot like $f(x) = 0$ for $x \ll 0$ and $f(x) = 1$ for $x \gg 0$. For a hypothesis $h_{\mathbf{w}}$, we can view $h_{\mathbf{w}}(\mathbf{x})$ as the probability that the label of $\mathbf{x}$ is 1.

We want to define a loss function which should determine how bad it is to predict $h_{\mathbf{w}}(\mathbf{x}) \in [0, 1]$ given that the true label is $y = \pm 1$. Suppose we have a sample (x, y) then our model outputs $\mathbb{P}(y = 1 \mid x)$, and for a given weight $\mathbf{w}$, this is $\mathbb{P}(y = 1 \mid x) = \sigma(\mathbf{w}^\top \mathbf{x})$. For a given set of independent samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we want to maximize

$$\mathcal{L}(\mathbf{w}) = \prod_{y_i=1} \mathbb{P}(y = 1 \mid \mathbf{x}_i) \cdot \prod_{y_i=0} \mathbb{P}(y = 0 \mid \mathbf{x}_i)$$

Since $x^0 = x$ for all x, we can write this as

$$= \prod_{i=1}^{n} \mathbb{P}(y = 1 \mid \mathbf{x}_i)^{y_i} \cdot \prod_{i=1}^{n} \mathbb{P}(y = 0 \mid \mathbf{x}_i)^{1-y_i} = \prod_{i=1}^{n} \sigma(\mathbf{w}^\top \mathbf{x}_i)^{y_i} \cdot \prod_{i=1}^{n} \big(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)\big)^{1-y_i}$$

Maximizing this is the same as minimizing $-\log \mathcal{L}(\mathbf{w})$ (because we can then use gradient descent to find the minimum). And so our total loss function is

$$Err(\mathbf{w}) = -\log \mathcal{L}(\mathbf{w}) = \sum_{i=1}^{n} y_i \log \sigma(\mathbf{w}^\top \mathbf{x}_i) + \sum_{i=1}^{n} (1 - y_i) \log\big(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i)\big)$$

Let us write σ for $\sigma(\mathbf{w}^\top \mathbf{x}_i)$, then computing the gradient gives

$$\nabla Err(\mathbf{w}) = -\sum_{i=1}^{n} \left(y_i \frac{1}{\sigma}(1 - \sigma)\mathbf{x}_i + (1 - y_i)\frac{-1}{1 - \sigma}\sigma(1 - \sigma)\mathbf{x}_i \right) = -\sum_{i=1}^{n} (y_i(1 - \sigma)\mathbf{x}_i - (1 - y_i)\sigma \mathbf{x}_i)$$

And so we get that

$$\nabla Err(\mathbf{w}) = -\sum_{i=1}^{n} \big(y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)\big)\mathbf{x}_i$$

Thus using pure stochastic gradient descent, we have the update rule

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{x}_i \big(\sigma(\mathbf{w}^\top \mathbf{x}_i) - y_i\big)$$

5 Multiclass Classification

5.1 Softmax Regression

Suppose we have K different classes we want to classify our input into (e.g. airplane, automobile, bird, cat, etc.). One way to do this is to define K binary classifiers and then return the class identified as true. Though this will of course not work if two binary classifiers both return true, and furthermore this is expensive.

Instead what we can do is have our output space be probability tuples, i.e. it is the K-simplex $\Delta^K = \{(p_1, \dots, p_K) \mid p_i \geq 0, \sum_i p_i = 1\}$. Suppose we have an input vector $\mathbf{x}$, then we can multiply it with weights $\mathbf{w}_1, \dots, \mathbf{w}_m$ to get $\mathbf{w}_1^\top \mathbf{x}, \dots, \mathbf{w}_m^\top \mathbf{x}$. Since the sum of these outputs is not necessarily 1, we can normalize these and return

$$p_i = \frac{\mathbf{w}_i^\top \mathbf{x}}{\sum_j \mathbf{w}_j^\top \mathbf{x}}$$

But this is not good enough; what if $\mathbf{w}_i^\top \mathbf{x}$ isn't positive? We will use the *softmax* function: a continuous function which approximates the maximum of a vector:

Definition 5.1.1

We define the **softmax** function to be a function softmax: $\mathbb{R}^n \longrightarrow \mathbb{R}^n$ with components $\text{softmax}_i \colon \mathbb{R}^n \longrightarrow \mathbb{R}$:

$$\text{softmax}_i(\mathbf{z}) = \frac{\exp(\mathbf{z}_i)}{\sum_j \exp(\mathbf{z}_j)}$$

we can then define the softmax of $\mathbf{z}$ to be $\text{softmax}(\mathbf{z}) = (\text{softmax}_1(\mathbf{z}), \dots, \text{softmax}_n(\mathbf{z}))$.

Notice that trivially the sum of $\text{softmax}_i(\mathbf{z})$ is 1, and $\text{softmax}_i(\mathbf{z}) > 0$ for all i. Thus we will define

$$p_i = \text{softmax}_i(\mathbf{w}_i^\top \mathbf{x}) = \frac{\exp\left(\mathbf{w}_i^\top \mathbf{x}\right)}{\sum_j \exp\left(\mathbf{w}_j^\top \mathbf{x}\right)}$$

In matrix form

$$\mathbf{p} = \text{softmax}(\mathbf{W}\mathbf{x})$$

where the rows of $\mathbf{W}$ are the weight vectors $\mathbf{w}_i$.

p_i is the probability that the class of $\mathbf{x}$ is i. We will also write it as $p_i = \mathbb{P}(y = i \mid \mathbf{x})$, i.e. the probability that the output class is i given the input x. If the variable y is in use, I will also use $\mathbb{P}(o = i \mid \mathbf{x})$ (o for output).

Definition 5.1.2

Suppose we are given a sequence of independent identically distributed training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Then the **likelihood** of a weight matrix $\mathbf{W}$ is defined to be

$$\mathcal{L}(\mathbf{W}) = \prod_{i=1}^n \prod_{k=1}^K \mathbb{P}(o_i = k \mid \mathbf{x}_i)^{\mathbf{1}\{y_i = k\}} = \prod_{i=1}^n \mathbb{P}(o_i = y_i \mid \mathbf{x}_i)$$

Then the **log-likelihood** is

$$\log \mathcal{L}(\mathbf{W}) = \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}\{y_i = k\} \log \mathbb{P}(o_i = k \mid \mathbf{x}_i) = \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}\{y_i = k\} \log \frac{\exp(\mathbf{w}_k^\top \mathbf{x}_i)}{\sum_j \exp(\mathbf{w}_j^\top \mathbf{x}_i)}$$

We can then perform stochastic gradient descent where the loss function is the log-likelihood.

5.2 Support Vector Machines (SVM)

Suppose we want to solve the problem of binary classification, i.e. we want to split our space into halfspaces. There can be many halfspaces which do so (think of the simplest problem, where we have two points), so we want to find some optimal halfspace. Let us define the *margin* of a class to a half space to be the minimum distance between the dividing hyperplane and a point from the class.

We need a quick result from linear algebra:

Proposition 5.2.1

Let $\mathcal{H}$ be the space defined by $\mathbf{w}^\perp$ for $\mathbf{w} \neq 0$. Show that for any $\mathbf{x}$, $\mathbf{x} - \pi_{\mathcal{H}}(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{w}\rangle}{\langle \mathbf{w}, \mathbf{w}\rangle}\mathbf{w}$.

Proof: First let us assume that $\mathbf{w}$ is normalized, then let us extend $\mathbf{w}$ to an orthonormal basis, $\{\mathbf{w}, \mathbf{w}_1, \dots, \mathbf{w}_n\}$. Then $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ is an orthonormal basis for $\mathcal{H}$. Now suppose that $\mathbf{x} = \alpha\mathbf{w} + \sum_i \alpha_i \mathbf{w}_i$, then $\pi_{\mathcal{H}}(\mathbf{x}) = \sum_i \alpha_i \mathbf{w}_i$ and so $\mathbf{x} - \pi_{\mathcal{H}}(\mathbf{x}) = \alpha\mathbf{w}$. And $\langle \mathbf{x}, \mathbf{w}\rangle = \alpha$, so we have that $\mathbf{x} - \pi_{\mathcal{H}}(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}\rangle\mathbf{w}$ as required (since $\mathbf{w}$'s norm is 1).

Now let $\widehat{\mathbf{w}}$ be the normalization of $\mathbf{w}$, i.e. $\widehat{\mathbf{w}} = \mathbf{w}/\|\mathbf{w}\|$. Then by above

$$\mathbf{x} - \pi_{\mathcal{H}}(\mathbf{x}) = \langle \mathbf{x}, \widehat{\mathbf{w}}\rangle\widehat{\mathbf{w}} = \frac{\langle \mathbf{x}, \mathbf{w}\rangle}{\langle \mathbf{w}, \mathbf{w}\rangle}\mathbf{w}$$

as required. ∎

If our hyperplane is given by $\mathcal{H} = \{\mathbf{x} \mid \langle \mathbf{w}, \mathbf{x}\rangle = 0\} = \mathbf{w}^\perp$ then the distance between a point $\mathbf{x}$ and $\mathcal{H}$ is just $\|\mathbf{x} - \pi_{\mathcal{H}}(\mathbf{x})\|$, and by the above proposition this is just

$$\left\| \frac{\langle \mathbf{x}, \mathbf{w}\rangle}{\langle \mathbf{w}, \mathbf{w}\rangle}\mathbf{w} \right\| = \frac{|\mathbf{w}^\top \mathbf{x}|}{\|\mathbf{w}\|}$$

So the margin of a class $\mathcal{C}$ is just

$$\min_{\mathbf{x}\in\mathcal{C}} \frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|}$$

In our situation we have two classes: those $\mathbf{x}$s where $\mathbf{w}^\top\mathbf{x} > 0$ and those where $\mathbf{w}^\top\mathbf{x} < 0$. So the margin of the first class is just $\min_{\mathbf{w}^\top\mathbf{x}>0} \frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|}$, and the margin of the second class is $\min_{\mathbf{w}^\top\mathbf{x}<0} -\frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|} = -\max_{\mathbf{w}^\top\mathbf{x}<0} \frac{\mathbf{w}^\top\mathbf{x}}{\|\mathbf{w}\|}$. That means that $\min_{\mathbf{w}^\top\mathbf{x}>0} \mathbf{w}^\top\mathbf{x} = -\max_{\mathbf{w}^\top\mathbf{x}<0} \mathbf{w}^\top\mathbf{x} = \varepsilon$.

So $y_i = 1$ ($\mathbf{x}_i$ is in the positive class) if and only if $\mathbf{w}^\top\mathbf{x}_i \geq \varepsilon$ and $y_i = -1$ if and only if $\mathbf{w}^\top\mathbf{x}_i \leq -\varepsilon$. That is we have that $y_i\mathbf{w}^\top\mathbf{x}_i \geq \varepsilon$. These form two new parallel hyperplanes, and their distance is $\frac{2\varepsilon}{\|\mathbf{w}\|}$. So to maximize the distance between these hyperplanes, we want to minimize $\|\mathbf{w}\|$, equivalently $\|\mathbf{w}\|^2$. Note that if $\mathbf{w}$ is an optimal solution so is $\alpha\mathbf{w}$ (since they define the same hyperplanes), so we can scale $\mathbf{w}$ so that $\varepsilon = 1$. Meaning we want to solve the problem of minimizing $\|\mathbf{w}\|^2$ with respect to the constraints $y_i\mathbf{w}^\top\mathbf{x}_i \geq 1$ for every sample $\mathbf{x}_i$.

5.3 Constrained Optimization

Suppose we have a problem, like in the previous subquestion, where we want to minimize some function $f(\mathbf{x})$ with respect to some constraints $g_i(\mathbf{x}) \leq 0$:

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \forall i\colon g_i(\mathbf{x}) \leq 0$$

This is equivalent to minimizing the function $J(\mathbf{x}) = f(\mathbf{x}) + \sum_i \text{penalty}(g_i(\mathbf{x}))$ where penalty $= I$ is the infinite step function: $I(x) = \infty$ if $x > 0$ and zero otherwise. But this is hard to minimize since penalty here is not continuous. Instead we can approximate I linearly, using the *Lagrangian*:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_i \lambda_i g_i(\mathbf{x})$$

Proposition 5.3.1

For every $\mathbf{x}$, $\max_{\boldsymbol{\lambda}\geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = J(\mathbf{x})$.

Proof: if $f_i(\mathbf{x}) \leq 0$ for all i then $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is maximized when $\boldsymbol{\lambda} = 0$ and then $\mathcal{L}(\mathbf{x}, 0) = f(\mathbf{x}) = J(\mathbf{x})$. If $f_i(\mathbf{x}) > 0$ for some i then taking $\lambda_i \to \infty$ we get $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \to \infty$, so the maximum is ∞, and $J(\mathbf{x}) = \infty$. ∎

Thus the problem is equivalent to $\min_{\mathbf{x}} \max_{\boldsymbol{\lambda}\geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$. Suppose we flip the minimum and maximum: $\max_{\boldsymbol{\lambda}\geq 0} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$. Let us define

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}), \qquad h(\mathbf{x}) = \max_{\boldsymbol{\lambda}\geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$$

Notice then that for all $\mathbf{x}, \boldsymbol{\lambda}$:

$$g(\boldsymbol{\lambda}) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \leq h(\mathbf{x})$$

Thus we can maximize $\boldsymbol{\lambda}$ and minimize $\mathbf{x}$ and we get

$$\max_{\boldsymbol{\lambda}\geq 0} g(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\geq 0} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) \leq \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}\geq 0} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} h(\mathbf{x})$$

Definition 5.3.2

The **Lagrangian dual program** of the primal program

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \forall i\colon g_i(\mathbf{x}) \leq 0$$

is the program of maximizing $g(\boldsymbol{\lambda})$: $\max_{\boldsymbol{\lambda}\geq 0} g(\boldsymbol{\lambda})$ where $g(\boldsymbol{\lambda}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$.

We have just shown

Theorem 5.3.3 (Weak Duality Principle)

The Lagrangian dual program gives a lower bound to the optimal solution of the primal program.

5.4 Solving Support Vector Machines

Recall that an SVM is of the form (we minimize $\frac{1}{2}\|\mathbf{w}\|^2$ so the gradient is nicer):

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i\colon y_i \mathbf{w}^\top \mathbf{x}_i \geq 1$$

which can be equivalently written as

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{s.t.} \quad \forall i\colon 1 - y_i \mathbf{w}^\top \mathbf{x}_i \leq 0$$

The Lagrangian is then

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \lambda_i (1 - y_i \mathbf{w}^\top \mathbf{x}_i)$$

Taking the gradient only with respect to $\mathbf{w}$ gives

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = \mathbf{w} - \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i$$

This is zero when

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}_i$$

Notice then that

$$\|\mathbf{w}\|^2 = \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

So plugging this back into the Lagrangian we get

$$g(\boldsymbol{\lambda}) = \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) = \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{n} \lambda_i - \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

Taking the partial derivative of $g(\boldsymbol{\lambda})$ with respect to λ_i gives

$$\frac{\partial}{\partial \lambda_i} g(\boldsymbol{\lambda}) = 1 - \frac{1}{2} \sum_{j=1}^{n} \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

So we can use this to compute the gradient ascent of (since we want to increase) $g(\boldsymbol{\lambda})$.

5.5 Solving Noise

Suppose our data is not separable; there is noise. That means that there is no halfspace which correctly classifies our training data. For each $\mathbf{x}_i$ we define a parameter ξ_i which will measure how far away $\mathbf{x}_i$ is from the correct class (the margin).

If we have that $y_i\mathbf{w}^\top\mathbf{x}_i \geq 1$ then $\mathbf{x}_i$ is properly classified and so ξ_i should be zero. Otherwise we want to measure how far away $\mathbf{x}_i$ is from the margin $\mathbf{w}^\top\mathbf{x} = 1$. The distance to the halfspace $\mathbf{w}^\top\mathbf{x} = 0$ is $\frac{|\mathbf{w}^\top\mathbf{x}_i|}{\|\mathbf{w}\|}$, and the distance from $\mathbf{w}^\top\mathbf{x} = 1$ to $\mathbf{w}^\top\mathbf{x} = 0$ is $\frac{1}{\|\mathbf{w}\|}$. If $\mathbf{x}_i$ is misclassified, then the first distance is $-\frac{y_i\mathbf{w}^\top\mathbf{x}_i}{\|\mathbf{w}\|}$, so the total distance to the correct margin is $\frac{1}{\|\mathbf{w}\|}(1 - y_i\mathbf{w}^\top\mathbf{x}_i)$. And if $\mathbf{x}_i$ is classified correctly but not beyond the margin, then we want the difference between the two distances, which is the same. So all in all we should have something like $\xi_i = \frac{1}{\|\mathbf{w}\|}(1 - y_i\mathbf{w}^\top\mathbf{x}_i)$. $\|\mathbf{w}\|$ is just a constant so we can simply set $\xi_i = 1 - y_i\mathbf{w}^\top\mathbf{x}_i$.

Notice that in the second case we have that $y_i\mathbf{w}^\top\mathbf{x}_i < 1$ and so $1 - y_i\mathbf{w}^\top\mathbf{x}_i > 0$ so we can just define

$$\xi_i = \max\{0, 1 - y_i\mathbf{w}^\top\mathbf{x}_i\}$$

So to solve noise what we can do is minimize $\frac{1}{2}\|\mathbf{w}\|^2 + c\sum_i \xi_i$ where c is the cost of a misclassification, according to the constraints $y_i\mathbf{w}^\top\mathbf{x}_i \geq 1 - \xi_i$.

5.6 Nonlinear SVMs

Suppose our data is not linearly separable: perhaps it is a circle. The idea to solve this is we can perform a transformation of the input space into some larger dimension, and hope that the inputs are now linearly classifiable in this larger dimension. So we define a *feature map* which is a function from the input space to a feature space, $\varphi\colon \mathbb{R}^d \longrightarrow \mathbb{R}^D$ where $D \gg d$. Recall that the gradient of $g(\boldsymbol{\lambda})$ has components of the form

$$\frac{\partial}{\partial\lambda_i} g(\boldsymbol{\lambda}) = 1 - \frac{1}{2}\sum_{j=1}^{n} \lambda_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

So if we now replace $\mathbf{x}_i$ with the feature $\varphi(\mathbf{x}_i)$, we get

$$\frac{\partial}{\partial\lambda_i} g(\boldsymbol{\lambda}) = 1 - \frac{1}{2}\sum_{j=1}^{n} \lambda_j y_i y_j \varphi(\mathbf{x}_i)^\top \varphi(\mathbf{x}_j)$$

Notice that φ always appears in an inner product, so we can define the *kernel* to be $k(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x})^\top\varphi(\mathbf{y})$, and then we have that

$$\frac{\partial}{\partial\lambda_i} g(\boldsymbol{\lambda}) = 1 - \frac{1}{2}\sum_{j=1}^{n} \lambda_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

For some types of feature maps computing k is simpler than φ. For example:

$$\varphi(\mathbf{x}) = \varphi\begin{pmatrix} x1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ x_1 \\ x_2 \\ 1 \end{pmatrix}$$

Then a computation gives us

$$k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top\mathbf{z} + 1)^2$$

So we can focus on kernels instead of feature maps, in fact we have the following result due to Moore-Aronszajn:

Theorem 5.6.1 (Moore-Aronszajn)

Every positive semi-definite function k defines a valid kernel.

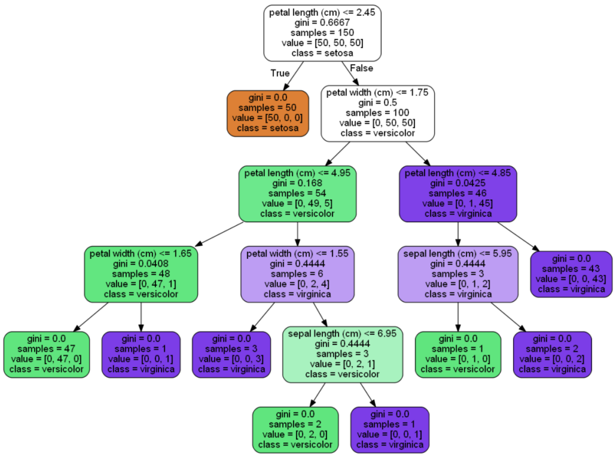
So for example, we can define

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{\sigma^2}\right)$$

and this is positive semi-definite so it defines a kernel.

5.7 Decision Trees

A *decision tree* is an iterative procedure where at each iteration data is split according to some cutoff on a certain feature. The features and cutoff points are chosen to minimize either variance: $\frac{1}{n}\sum_y (y - \bar{y})^2$ (for numeric/real y), or entropy: $-\sum_y p(y)\log p(y)$ (for discrete y) where $p(y)$ is the probability of a sample belong to class y. Take for example the following decision tree to classify flowers:



Each node has the following information:

(1) The feature and cutoff: the features here are petal width and sepal length, and the cutoffs are the conditions at the top of each (non-leaf) node.

(2) Gini: an alternative for entropy and variance, ignore this.

(3) The number of samples: the number of samples remaining to be classified.

(4) The value: an array showing how many of the current samples belong to each class.

(5) The class: the current class the set of samples are believed to belong to.

Here we need to use entropy and not variance since the classes are discrete (and not numeric). The entropy of the first node is $-\frac{1}{3}\log\frac{1}{3}-\frac{1}{3}\log\frac{1}{3}-\frac{1}{3}\log\frac{1}{3}=\log 3$, since the probability of being in a class is $\frac{1}{3}$ (by the values array). The entropy of its left child is $-1\log 1-0\log 0-0\log 0=0$ (since there is no unknown, the entropy should be zero). The entropy of its right child is $-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}=1$.

5.8 Boosting and Gradient Boosting

Boosting is an ensemble-based approach to machine learning (meaning it uses many other learning algorithms to improve performance). Suppose we have a labeled sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ which are independent and identically distributed. Furthermore we have a loss function $\ell(\bullet,\bullet)$. We want to learn a function f which minimizes the empirical risk (as this is an approximation to minimizing generalized risk):

$$\mathcal{L}(f)=\frac{1}{n}\sum_i \ell(y_i, f(\mathbf{x}_i))$$

But there are far too many functions to choose from, so instead we consider a class $\mathcal{H}$ of simple base learners (typically small decision trees), and focus on functions of the form

$$f(\mathbf{x})=\sum_i \gamma_i h_i(\mathbf{x}), \qquad h_i\in\mathcal{H}$$

Typically in gradient descent we have an update rule of the form $\theta_{t+1}=\theta_t-\gamma_t\frac{\partial}{\partial\theta}\mathcal{L}(\theta_t)$. Here our update rule will be of the form

$$f_{t+1}=f_t-\gamma_{t+1}\frac{\partial}{\partial f}\mathcal{L}(f_t)$$

Our iterations are of the form $f_t=\sum_{i=1}^t \gamma_i h_i$, and so $f_{t+1}=f_t+\gamma_{t+1}h_{t+1}$. Thus we want to learn $h_{t+1}\in\mathcal{H}$ to fit $-\frac{\partial}{\partial f}\mathcal{L}(f_t)$, so that f_{t+1} is an approximation to the step we got in gradient descent. And we choose γ_{t+1} to minimize the overall loss:

$$\gamma_{t+1}=\underset{\gamma}{\operatorname{argmin}}\,\mathcal{L}(f_t+\gamma h_{t+1})=\underset{\gamma}{\operatorname{argmin}}\sum_{i=1}^n \ell(y_i, f_t(\mathbf{x}_i)+\gamma h_{t+1}(\mathbf{x}_i))$$

The gradient boosting algorithm is as follows:

```
function GRADIENT BOOSTING($\{\mathbf{x}_i, y_i\}_{i=1}^n$)
```

```
    $f_1 \leftarrow c$ for some constant $c$
    for $(t = 1, \ldots, M)$
        compute the residuals: $r_i = \frac{\partial}{\partial f(\mathbf{x})} \ell(y_i, f_t(\mathbf{x}_i))$
        learn $h_{t+1} \in \mathcal{H}$ to the new training data $\{\mathbf{x}_i, r_i\}_{i=1}^{n}$
        find $\gamma_{t+1}$ which minimizes $\mathcal{L}(f_t + \gamma_{t+1} h_{t+1})$
        $f_{t+1} \leftarrow f_t + \gamma_{t+1} h_{t+1}$
    end for
end function
```

Some common loss functions include:

- Continuous y: MSE: $\ell(y, f(\mathbf{x})) = \frac{1}{2}(y - f(\mathbf{x}))^2$. Then the residual is $\frac{\partial}{\partial \ell}(y_i, f_t(\mathbf{x}_i)) = -(y_i - f_t(\mathbf{x}_i))$.
- Categorical/discrete y: consider $y = \pm 1$

  (**1**) Bernoulli loss: $\ell(y, f(\mathbf{x})) = \log(1 + \exp(-2yf(\mathbf{x})))$.

  (**2**) Adaboost loss: $\ell(y, f(\mathbf{x})) = \exp(-2yf(\mathbf{x}))$.

6 Unsupervised Learning

6.1 k-Means

Suppose we have samples $\{\mathbf{x}_i\}_{i=1}^n$ from $\mathbb{R}^d$ and a number of clusters K. We want to assign to every $\mathbf{x}_i$ a cluster in $\{1,\dots,K\}$. The idea behind our algorithm is to find a set of *prototypes* $\boldsymbol{\mu} = \{\boldsymbol{\mu}_i\}_{i=1}^n \subseteq \mathbb{R}^d$ and we cluster $\mathbf{x}_i$ according to which prototype it is closest to. That is, we want to find the j which minimizes $d(\mathbf{x}_i, \boldsymbol{\mu}) = \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$, and we assign to $\mathbf{x}_i$ the cluster j.

The problem is that we want to minimize $D(\boldsymbol{\mu}) = \sum_i d(\mathbf{x}_i, \boldsymbol{\mu})$, i.e. find $\boldsymbol{\mu}^* = \operatorname{argmin}_{\boldsymbol{\mu}} D(\boldsymbol{\mu})$. Unfortunately this problem is NP-hard, so instead we present an algorithm to iteratively improve $D(\boldsymbol{\mu})$.

```
function k-Means($\{\mathbf{x}_i\}_{i=1}^n$)
    let $\{\boldsymbol{\mu}_j^0\}_{j=1}^K$ be randomly initialized
    for ($t = 0, \dots, M$)
        for each $i$, cluster($\mathbf{x}_i$) $\leftarrow \operatorname{argmin}_j \|\mathbf{x}_i, \boldsymbol{\mu}_j^t\|^2$
        for each $j$, $\boldsymbol{\mu}_j^{t+1} = \operatorname{mean}\{\mathbf{x}_i \mid \text{cluster}(\mathbf{x}_i) = j\}$
    end for
    return $\{\boldsymbol{\mu}_j\}_{j=1}^K$
end function
```

Lemma 6.1.1

Let $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, then

$$\operatorname{mean}\{\mathbf{x}_i\}_{i=1}^n = \operatorname{argmin} \mathcal{L}(\boldsymbol{\mu}) = \operatorname{argmin} \sum_i \|\mathbf{x}_i - \boldsymbol{\mu}\|^2$$

Proof: we differentiate $\mathcal{L}(\boldsymbol{\mu})$ and get

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L} = 2 \sum_i (\mathbf{x}_i - \boldsymbol{\mu})$$

This is zero if and only if

$$\boldsymbol{\mu} = \frac{1}{n} \sum_i \mathbf{x}_i = \operatorname{mean}\{\mathbf{x}_i\}_{i=1}^n \qquad \blacksquare$$

Proposition 6.1.2

In the above algorithm, $D(\boldsymbol{\mu})$ improves every iteration.

Proof: let $\boldsymbol{\mu}^t = \{\boldsymbol{\mu}_j^t\}_{j=1}^K$, so we want to show that $D(\boldsymbol{\mu}^t) \geq D(\boldsymbol{\mu}^{t+1})$. Let us define $r_{ij} = 1$ if $\mathbf{x}_i$ is in cluster j and zero otherwise. Notice then that

$$d(\mathbf{x}_i, \boldsymbol{\mu}) = \min_j \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 = \min_{r_{ij}} \sum_j r_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

Thus

$$\min_{\boldsymbol{\mu}} D(\boldsymbol{\mu}) = \min_{\boldsymbol{\mu}} \sum_i d(\mathbf{x}_i, \boldsymbol{\mu}) = \min_{\boldsymbol{\mu}} \min_{r_{ij}} \sum_{i,j} r_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$$

let us denote this as $\min_{\boldsymbol{\mu}} \min_{\mathbf{r}} g(\boldsymbol{\mu}, \mathbf{r})$, so $D(\boldsymbol{\mu}) = \min_{\mathbf{r}} g(\boldsymbol{\mu}, \mathbf{r})$. Now let us define a new procedure where every iteration takes two steps:

(1) $\mathbf{r}^{t+1} = \operatorname{argmin}_{\mathbf{r}} g(\boldsymbol{\mu}^t, \mathbf{r})$

(2) $\boldsymbol{\mu}^{t+1} = \operatorname{argmin}_{\boldsymbol{\mu}} g(\boldsymbol{\mu}, \mathbf{r}^{t+1})$

After these two steps we have that

$$g(\boldsymbol{\mu}^{t+1}, \mathbf{r}^{t+1}) = \min_{\mathbf{r}} g(\boldsymbol{\mu}^{t+1}, \mathbf{r}) = D(\boldsymbol{\mu}^{t+1})$$

And since $g(\boldsymbol{\mu}^t, \mathbf{r}^t) \geq g(\boldsymbol{\mu}^{t+1}, \mathbf{r}^{t+1})$ we have that $D(\boldsymbol{\mu}^{t+1}) \leq D(\boldsymbol{\mu}^t)$.

Now we will show that this procedure is equivalent to k-Means.

(1) $\mathbf{r}^{t+1} = \operatorname{argmin}_{\mathbf{r}} g(\boldsymbol{\mu}^t, \mathbf{r})$: this minimizes $\min_{\mathbf{r}} \sum_j r_{ij} \left\|\mathbf{x}_i - \boldsymbol{\mu}_j^t\right\|^2$, i.e. we find assignments r_{ij} which minimizes $\left\|\mathbf{x}_i - \boldsymbol{\mu}_j^t\right\|^2$. This is what line 4 does in k-MEANS.

(2) $\boldsymbol{\mu}^{t+1} = \operatorname{argmin}_{\boldsymbol{\mu}} g(\boldsymbol{\mu}, \mathbf{r}^{t+1})$, this finds $\boldsymbol{\mu}_j$ which minimizes $\min_{\boldsymbol{\mu}_j} \sum_{\text{cluster}(\mathbf{x}_i)=j} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$. By the above lemma, this just means $\boldsymbol{\mu}_j = \text{mean}\{\mathbf{x}_i \mid \text{cluster}(\mathbf{x}_i) = j\}$, which is what line 5 does.

■

6.2 Expectation Maximization

Recall the following result:

Theorem 6.2.1 (Jensen's Inequality)

Let X be a random variable and $\varphi\colon \mathbb{R} \longrightarrow \mathbb{R}$ a convex function then

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi X]$$

Definition 6.2.2

Let P, Q be two distributions on $\mathcal{X}$ with pdfs p, q respectively. Then their **Kullback-Leibler divergence** (KL divergence) is defined as

$$D_{\text{KL}}(P\|Q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\, dx = \mathbb{E}_{x\sim P}\left[\log \frac{p(x)}{q(x)}\right]$$

Theorem 6.2.3

$D_{\text{KL}}(P\|Q) \geq 0$

Proof: by flipping the fraction, we have

$$D_{\text{KL}}(P\|Q) = \mathbb{E}_{x\sim P}\left[-\log \frac{q(x)}{p(x)}\right] \geq -\log \mathbb{E}_{x\sim P}\left[\frac{q(x)}{p(x)}\right]$$

where the inequality is due to Jensen. Continuing, we have

$$= -\log \int_{\mathcal{X}} p(x) \frac{q(x)}{p(x)}\, dx = -\log \int_{\mathcal{X}} q(x) = -\log 1 = 0$$

■

Suppose we have samples $\mathbf{X} \subseteq \mathcal{X}$ whose distribution is dependent on some latent variables $\mathbf{Z} \subseteq \mathcal{Z}$. Suppose we know the joint distribution of $(\mathbf{X}, \mathbf{Z})$ up to some unkown $\boldsymbol{\theta}$, $(\mathbf{X}, \mathbf{Z}) \sim P(\boldsymbol{\theta})$, and let the distribution of $\mathbf{Z}$ be Q. Let the pdf of Q be q and the pdf of $P(\boldsymbol{\theta})$ be $p_{\boldsymbol{\theta}}$. Given that we've seen $\mathbf{X}$ we want to find the $\boldsymbol{\theta}$ which maximizes $\mathbb{P}(\boldsymbol{\theta} \mid \mathbf{X})$, this can be written as

$$\mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta}) = \int_{\mathcal{Z}} \mathbb{P}(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})\, d\mathbf{Z} = \int_{\mathcal{Z}} p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z})\, d\mathbf{Z}$$

This is unfortunately hard to maximize.

So instead let us focus on maximizing $\mathbb{E}[\log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})]$. We know that $(\mathbf{X}, \mathbf{Z}) \sim P(\boldsymbol{\theta})$ and $\mathbf{Z} \sim Q$ so

$$\mathbb{E}[\log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})] = \int_{\mathcal{Z}} q(\mathbf{Z}) \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta}, \mathbf{Z})\, d\mathbf{Z} = \int_{\mathcal{Z}} q(\mathbf{Z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\, d\mathbf{Z}$$

Let us denote this

Definition 6.2.4

Define the **elbo** function as

$$\mathcal{L}(q,\boldsymbol{\theta}) = \mathbb{E}[\log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})] = \int_{\mathcal{Z}} q(\mathbf{Z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\, d\mathbf{Z}$$

Notice that by Jensen

$$\mathcal{L}(q,\boldsymbol{\theta}) \le \log \mathbb{E}[\mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})] = \log \int_{\mathcal{Z}} q(\mathbf{Z})\, \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta}, \mathbf{Z})\, d\mathbf{Z} = \log \int_{\mathcal{Z}} p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})\, d\mathbf{Z} = \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$$

So the elbo function is a lower bound for $\log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$, thus maximizing it should give us an approximation for the maximum of $\log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$.

Also notice that

$$\begin{aligned}
\mathcal{L}(q,\boldsymbol{\theta}) &= \int_{\mathcal{Z}} q(\mathbf{Z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\, d\mathbf{Z} = \int_{\mathcal{Z}} q(\mathbf{Z}) \log \frac{\mathbb{P}(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})\, \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})}{q(\mathbf{Z})}\, d\mathbf{Z} \\
&= \int_{\mathcal{Z}} q(\mathbf{Z}) \log \frac{\mathbb{P}(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})}{q(\mathbf{Z})}\, d\mathbf{Z} + \int_{\mathcal{Z}} q(\mathbf{Z}) \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})\, d\mathbf{Z} \\
&= - D_{\mathrm{KL}}(\mathbf{Z}, (\mathbf{Z} \,|\, \mathbf{X},\boldsymbol{\theta})) + \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})
\end{aligned}$$

Similarly we have

$$\mathcal{L}(q,\boldsymbol{\theta}) = \int_{\mathcal{Z}} q(\mathbf{Z}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\, d\mathbf{Z} = \int_{\mathcal{Z}} q(\mathbf{Z}) \log\bigl(p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})\bigr) - \int_{\mathcal{Z}} q(\mathbf{Z}) \log q(\mathbf{Z}) = \int_{\mathcal{Z}} q(\mathbf{Z}) \log\bigl(p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})\bigr) - H(q)$$

The expectation-maximization (EM) algorithm works in two steps: the E step and the M step. First in the E-step we maximize $\mathcal{L}(q,\boldsymbol{\theta})$ with respect to q, holding $\boldsymbol{\theta}$ constant. In the M-step we maximize $\mathcal{L}(q,\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, holding q constant.

(1) E-step: we want to maximize $-D_{\mathrm{KL}}(\mathbf{Z}, (\mathbf{Z} \,|\, \mathbf{X},\boldsymbol{\theta})) + \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$ with respect to q. Notice that the second term is independent on q, and the first term is non-positive. So $\mathcal{L}(q,\boldsymbol{\theta})$ is maximized when the first term is zero, which is when $\mathbf{Z}$ distributes the same as $\mathbf{Z} \,|\, \mathbf{X},\boldsymbol{\theta}$, i.e. $q(\mathbf{Z}) = \mathbb{P}(\mathbf{Z} \mid \mathbf{X},\boldsymbol{\theta})$. And in such a case we have $\max_q \mathcal{L}(q,\boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$.

(2) M-step: we want to maximize $\int_{\mathcal{Z}} q(\mathbf{Z}) \log\bigl(p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})\bigr) - H(q)$ with respect to $\boldsymbol{\theta}$. Since $H(q)$ does not depend on $\boldsymbol{\theta}$, we want to maximize $\int_{\mathcal{Z}} q(\mathbf{Z}) \log p_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{Z})$. This can be easier than optimizing $\mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$ if $p_{\boldsymbol{\theta}}$ is a nice function.

Since at each E-step we have $\mathcal{L}(q,\boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$, and at each M-step we increase $\mathcal{L}(q,\boldsymbol{\theta})$, it follows that $\mathbb{P}(\mathbf{X} \mid \boldsymbol{\theta})$ is increased each iteration.

Example 6.2.5

Suppose we have two biased coins, one with a bias of θ_A and the other with a bias of θ_B. Meaning when we flip coin A, its result distributes $\mathrm{Ber}(\theta_A)$. Now suppose we get a sequence of flips, where the sequence is constructed as follows: we choose a coin randomly with equal probability, then flip it ten times, and repeat this process five times total. Can we estimate θ_A, θ_B? If we know which coin was flipped for each flip then this is simple (suppose coin A gave 24 heads and 6 tails, then θ_A can be approximated as $24/(24+6)$, this is the MLE estimator). But suppose we don't know which coin was flipped, then we have a latent variable Z which has a bernoulli distribution which determines which coin we flip.

Our space $\mathcal{X}$ consists of 5-tuples of 10-tuples of flips (i.e. $(\{0,1\}^{10})^5$), and $\mathcal{Z}$ is $\{A,B\}^5$. Suppose our flips are (ignoring the order):

Flip #	# Heads	# Tails
$\mathbf{X}_1$	5	5
$\mathbf{X}_2$	9	1
$\mathbf{X}_3$	8	2
$\mathbf{X}_4$	4	6
$\mathbf{X}_5$	7	3

So we want to maximize $\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p_\theta(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}$. We begin with an estimate for $\boldsymbol{\theta}$: $\boldsymbol{\theta}^0$ which we can choose randomly. Suppose we choose $\theta_A^0 = 0.6$ and $\theta_B^0 = 0.5$. Now we want to compute $\mathbb{P}(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^0)$ for the E-step.

Note that in the first row we have 5 heads and 5 tails, the probability of this occurring with a bias of $\theta_A^0 = 0.6$ is $0.6^5 \cdot 0.4^5 \approx 0.0008$ and the probability of this occurring with a bias of $\theta_B^0 = 0.5$ is $0.5^5 \cdot 0.5^5 \approx 0.001$. That is $\mathbb{P}(\mathbf{X}_1 \mid \mathbf{Z}_1 = A, \boldsymbol{\theta}^0) = 0.0008$ and $\mathbb{P}(\mathbf{X}_2 \mid \mathbf{Z}_1 = B, \boldsymbol{\theta}^0) = 0.001$. Now we know that $\mathbb{P}(\mathbf{Z}_1 = A) = \mathbb{P}(\mathbf{Z}_1 = B) = 0.5$ (and this is independent of $\boldsymbol{\theta}$) so we get that

$$\mathbb{P}(\mathbf{Z}_1 = A \mid \mathbf{X}_1, \boldsymbol{\theta}^0) = \mathbb{P}(\mathbf{X}_1 \mid \mathbf{Z}_1 = A, \boldsymbol{\theta}) \cdot \frac{\mathbb{P}(\mathbf{Z}_1 = A \mid \boldsymbol{\theta}^0)}{\mathbb{P}(\mathbf{X}_1 \mid \boldsymbol{\theta}^0)}$$

Now we know that

$$\mathbb{P}(\mathbf{X}_1 \mid \boldsymbol{\theta}^0) = \sum_{i=A,B} \mathbb{P}(\mathbf{X}_1 \mid \mathbf{Z}_1 = i, \boldsymbol{\theta}^0) \cdot \mathbb{P}(\mathbf{Z}_1 = i) = 0.5(0.0008 + 0.001)$$

So we get that

$$\mathbb{P}(\mathbf{Z}_1 = A \mid \mathbf{X}_1, \boldsymbol{\theta}^0) = \frac{0.0008 \cdot 0.5}{0.5(0.0008 + 0.001)} = \frac{0.0008}{0.0008 + 0.001} \approx 0.45$$

And similarly

$$\mathbb{P}(\mathbf{Z}_1 = B \mid \mathbf{X}_1, \boldsymbol{\theta}^0) = \frac{0.001}{0.0008 + 0.001} \approx 0.55$$

Continuing this computation we get

i	$\mathbb{P}(\mathbf{Z}_i = A \mid \mathbf{X}_i, \boldsymbol{\theta}^0)$	$\mathbb{P}(\mathbf{Z}_i = B \mid \mathbf{X}_i, \boldsymbol{\theta}^0)$
1	0.45	0.55
2	0.8	0.2
3	0.73	0.27
4	0.35	0.65
5	0.65	0.35

So what this tells us is that we can view 0.45 of the first ten flips to belong to coin A and 0.55 of the first ten flips to belong to coin B and so on. Since in the first ten flips we have 5 heads and 5 tails, $0.45 \cdot 5 \approx 2.2$ of them are from coin A, etc. So we get the following table

Flip #	Coin A		Coin B	
1	2.2 H	2.2 T	2.8 H	2.8 T
2	7.2 H	0.8 T	1.8 H	0.2 T
3	5.9 H	1.5 T	2.1 H	0.5 T
4	1.4 H	2.1 T	2.6 H	3.9 T
5	4.5 H	1.9 T	2.5 H	1.1 T
Total	21.3 H	8.6 T	11.7 H	8.4 T

For the M-step we will use the MLE on these values. We get that

$$\theta_A^1 = \frac{21.3}{21.3 + 8.6} \approx 0.71, \qquad \theta_B^1 = \frac{11.7}{11.7 + 8.4} \approx 0.58$$

And then we continue the next iteration with $\boldsymbol{\theta}^1$.

6.3 The Gaussian Mixture Model (GMM)

Suppose we have samples $\{\mathbf{x}_i\}_{i=1}^n$ which are sampled from K multivariate Gaussian distributions of dimension d. That is we have a latent variable $\mathbf{z}_i$ telling us which Gaussian distribution $\mathbf{x}_i$ is sampled from:

$$\mathbf{x}_i \mid (\mathbf{z}_i = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$$

and suppose $\mathbb{P}(\mathbf{z}_i = k) = \pi_k$, π_k are called the **mixing values**. Suppose that π_k, μ_k, Σ_k are all unknowns, so we have $\boldsymbol{\theta} = (\pi_k, \mu_k, \Sigma_k)_{k=1}^K$. Then we have

$$\mathbb{P}(\mathbf{x}_1, \dots, \mathbf{x}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^n \mathbb{P}(\mathbf{x}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^n \sum_{k=1}^K \mathbb{P}(\mathbf{x}_i \mid \boldsymbol{\theta}, \mathbf{z}_i = k)\, \mathbb{P}(\mathbf{z}_i = k) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)$$

Where $\mathcal{N}(\mathbf{x} \mid \mu, \Sigma)$ is the multivariate normal density:

$$\mathcal{N}(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu)\right)$$

So let us define

$$\ell(\boldsymbol{\theta}) = \log \mathbb{P}(\mathbf{x}_1, \dots, \mathbf{x}_n \mid \boldsymbol{\theta}) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)\right)$$

Unfortunately, this is hard to maximize (differentiate).

Instead let us redefine our latent variables $\mathbf{Z}_i = (z_{i,1}, \dots, z_{i,K})$ where $z_{i,k} = 1$ when $\mathbf{x}_i$ is chosen from $\mathcal{N}(\mu_k, \Sigma_k)$ and 0 otherwise (i.e. it indicates whether $\mathbf{x}_i$ comes from cluster k). Thus

$$\mathbb{P}(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{Z}_1, \dots, \mathbf{Z}_n \mid \boldsymbol{\theta}) = \prod_{i=1}^n \mathbb{P}(\mathbf{x}_i, \mathbf{Z}_i) = \prod_{i=1}^n \prod_{k=1}^K \pi_k^{z_{i,k}} \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)^{z_{i,k}}$$

Then the log-likelihood is

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{Z}_1, \dots, \mathbf{Z}_n) = \sum_{i=1}^n \sum_{k=1}^K z_{i,k}\big(\log \pi_k + \log \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)\big)$$

Now for the E-step we set $q(z_{i,k}) = \mathbb{P}(z_{i,k} \mid \mathbf{x}_i, \boldsymbol{\theta})$, which is

$$q(z_{i,k} = 1) = \mathbb{P}(z_{i,k} = 1 \mid \mathbf{x}_i, \boldsymbol{\theta}) = \frac{\mathbb{P}(z_{i,k} = 1, \mathbf{x}_i \mid \boldsymbol{\theta})}{\mathbb{P}(\mathbf{x}_i \mid \boldsymbol{\theta})} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)}$$

Recall that $z_{i,k} = 1$ means that $\mathbf{x}_i$ comes from cluster k, so this is the probability of that.

Now for the M-step we want to maximize

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{Z} \sim q}[\log \mathbb{P}(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})] = \sum_{i=1}^n \sum_{k=1}^K q(z_{i,k} = 1) \log \mathbb{P}(\mathbf{x}_i, z_{i,k} = 1 \mid \boldsymbol{\theta})$$

and recall that $\mathbb{P}(\mathbf{x}_i, z_{i,k} = 1) = \pi_k \mathcal{N}(\mathbf{x}_i \mid \mu_k, \Sigma_k)$ so

$$\log \mathbb{P}(\mathbf{x}_i, z_{i,k} = 1) = \log \pi_k + \log \frac{1}{\sqrt{(2\pi)^d \det \Sigma_k}} - \frac{1}{2}(\mathbf{x}_i - \mu_k)^\top \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)$$

We now want to differentiate $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, so we must differentiate it with respect to every element in $\boldsymbol{\theta}$. We start with μ_k:

$$\frac{\partial}{\partial \mu_k} \log \mathbb{P}(\mathbf{x}_i, z_{i,k} = 1 \mid \boldsymbol{\theta}) = \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)$$

Now we get that

$$\frac{\partial}{\partial \mu_k} \mathcal{L}(q, \boldsymbol{\theta}) = 0 \iff \sum_i \sum_k q(z_{i,k} = 1) \Sigma_k^{-1}(\mathbf{x}_i - \mu_k) = 0$$

This can occur when

$$\sum_i q(z_{i,k} = 1)(\mathbf{x}_i - \mu_k) = 0 \iff \mu_k^{\text{new}} = \frac{\sum_i q(z_{i,k} = 1)\mathbf{x}_i}{\sum_i q(z_{i,k} = 1)}$$

Similarly for Σ_k we have

$$\frac{\partial}{\partial \Sigma_k} \log \mathbb{P}(\mathbf{x}_i, z_{i,k} = 1 \mid \boldsymbol{\theta}) = (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top$$

So we get that by setting $\frac{\partial}{\partial \Sigma_k}\mathcal{L}(q, \boldsymbol{\theta}) = 0$ we have

$$\Sigma_k^{\text{new}} = \frac{\sum_i q(z_{i,k} = 1)(\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top}{\sum_i q(z_{i,k} = 1)}$$

And finally we want to maximize with respect to π_k under the constraint that $\sum_k \pi_k = 1$. Using Lagrangian multipliers, we get

$$\pi_k^{\text{new}} = \frac{\sum_i q(z_{i,k} = 1)}{n}$$

6.4 Principal Component Analysis (PCA)

Deep neural networks (which will be studied more later) split classification into two parts: learning a representation, and classification of the representation. Consider a baby who is not taught the name of objects; he will still be able to learn representations of these objects (know that object A and object B have the same representation), he just won't be able to classify them as their name (e.g. a cup).

PCA is a form of unsupervised learning which will learn a representation of information on a smaller dimension than the input. That is, PCA maps input to a lower dimension while (hopefully) preserving the important characteristics.

Let $(\mathbf{v}_1, \dots, \mathbf{v}_r) = V \in M_{d\times r}(\mathbb{R})$ be an orthonormal set, a basis for a subspace. Note then that if $\mathbf{x} \in \text{span}(V)$, $\mathbf{x} = \sum_i a_i v_i$ for some $a_i \in \mathbb{R}$, meaning $\mathbf{x} = V\mathbf{a}$ for $\mathbf{a} = (a_1, \dots, a_r)^\top$. Since $\mathbf{v}_1, \dots, \mathbf{v}_r$ are orthonormal, V is an orthonormal (but not square) matrix, so $V^\top V = I_r$ (but not necessarily $VV^\top = I_d$, unless $d = r$).

Now consider an encoder-decoder construct where we encode a vector $\mathbf{x}$ by $\mathbf{a} = V^\top \mathbf{x}$, and decode it by $V\mathbf{a}$. Take a vector $\mathbf{x}$, then notice that $VV^\top\mathbf{x}$ is the projection of $\mathbf{x}$ onto $\text{span}(V)$: suppose we extend to an orthonormal basis $(\mathbf{v}_1, \dots, \mathbf{v}_d)$ then $\mathbf{x} = \sum_i a_i\mathbf{v}_i$ and $V^\top\mathbf{x} = \sum_{i=1}^r a_i\mathbf{e}_i$ and thus $VV^\top\mathbf{x} = \sum_{i=1}^r a_i\mathbf{v}_i = \pi_V(\mathbf{x})$. So $VV^\top\mathbf{x} = \mathbf{x}$ if and only if $\mathbf{x} \in \text{span}(V)$.

So our goal is to find the orthonormal basis V which distorts our samples $\{\mathbf{x}_i\}_{i=1}^n$ as little as possible. We know that V is orthonormal if and only if $V^\top V = I_r$, so we want to find V such that $V^\top V = I$ with the smallest total distortion:

$$\underset{V^\top V = I}{\text{argmin}} \sum_i \left\| VV^\top \mathbf{x}_i - \mathbf{x}_i \right\|^2$$

Now, we will assume that the mean of $\{\mathbf{x}_i\}_{i=1}^n$ is zero, so we must map $\mathbf{x}_i \mapsto \mathbf{x}_i - \bar{\mathbf{x}}$ where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$. This will be explained later.

Let us define $\mathbf{X}$ to be the matrix whose rows are $\mathbf{x}_i$, that is

$$\mathbf{X} = \begin{pmatrix} \rule{1cm}{0.4pt} & \mathbf{x}_1 & \rule{1cm}{0.4pt} \\ & \vdots & \\ \rule{1cm}{0.4pt} & \mathbf{x}_n & \rule{1cm}{0.4pt} \end{pmatrix} \in M_{n\times d}(\mathbb{R})$$

then we define the *covariance matrix* of $\mathbf{X}$ to be

$$\text{Cov}(\mathbf{X}) = \frac{1}{n}\mathbf{X}^\top\mathbf{X} \in M_d(\mathbb{R})$$

this approximates the covariance matrix of the distribution of the samples, that is it approximates $\text{Cov}(\mathcal{D}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\right]$ (here it is $\mathbf{x}\mathbf{x}^\top$ and not $\mathbf{x}^\top\mathbf{x}$ since $\mathbf{x}$ is a column vector). Assuming the mean is zero, the covariance matrix quantifies how much each sample/feature correlates with the other samples/features.

Note that $\text{Cov}(\mathbf{X})$ is a symmetric positive-definite matrix. This means that its eigenvalues are nonnegative real, and its eigenvectors can be taken to be orthonormal. Thus it has a *spectral decomposition*:

$$\text{Cov}(\mathbf{X}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$$

where $\boldsymbol{U}$ is the orthogonal matrix whose columns are the eigenvectors of $\text{Cov}(\mathbf{X})$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ are the eigenvalues ordered such that $\lambda_1 \geq \cdots \geq \lambda_n$. Let us write the associated eigenvector as $\mathbf{u}_1, \dots, \mathbf{u}_n$ (i.e. the eigenvector whose eigenvalue is λ_i is $\mathbf{u}_i$), so U's columns are (in order) $\mathbf{u}_i$.

Theorem 6.4.1 (PCA Solution)

$$\underset{V^\top V = I}{\operatorname{argmin}} \sum_i \left\| V V^\top \mathbf{x}_i - \mathbf{x}_i \right\|^2$$

where $\mathbf{v}_i = \mathbf{u}_i$ for $1 \leq i \leq r$. That is, the columns of V are the first r eigenvectors of $\operatorname{Cov}(\mathbf{X})$.

Let $Y = VV^\top \mathbf{X}^\top - X^\top$, then the columns of Y are $VV^\top \mathbf{x}_i - \mathbf{x}_i$. This means that the diagonals of $Y^\top Y$ are $\left\| VV^\top \mathbf{x}_i - \mathbf{x}_i \right\|^2$, and so the total distortion is $\operatorname{trace}\left(Y^\top Y\right) = \operatorname{trace}\left(YY^\top\right)$. Then using the fact that we can commute matrices inside the trace, we have

$$\operatorname{trace}\left(YY^\top\right) = \operatorname{trace}\left(VV^\top \mathbf{X}^\top \mathbf{X} V V^\top\right) - 2\operatorname{trace}\left(V^\top \mathbf{X}^\top \mathbf{X} V\right) + \operatorname{trace}\left(\mathbf{X}\mathbf{X}^\top\right)$$

Now we know that $V = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$ so $V^\top \operatorname{Cov}(\mathbf{X}) V = \operatorname{diag}(\lambda_1, \ldots, \lambda_r) = \mathbf{\Lambda}_r$. Since $\operatorname{Cov}(\mathbf{X}) = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$, this means that $V^\top \mathbf{X}^\top \mathbf{X} V = n\mathbf{\Lambda}_r$. So we have

$$= n \operatorname{trace}\left(V \mathbf{\Lambda}_r V^\top\right) - 2n \operatorname{trace}(\mathbf{\Lambda}_r) + \sum_{i=1}^{n} \|\mathbf{x}_i\|^2$$

since we can commute inside the trace and $V^\top V = I$, this is equal to

$$\sum_{i=1}^{n} \|\mathbf{x}_i\|^2 - n \sum_{i=1}^{r} \lambda_i$$

We know that $\operatorname{trace}(\operatorname{Cov}(\mathbf{X})) = \sum_{i=1}^{d} \lambda_i$ and so $\operatorname{trace}\left(\mathbf{X}\mathbf{X}^\top\right) = n\sum_{i=1}^{d} \lambda_i$. And we know that $\operatorname{trace}\left(\mathbf{X}\mathbf{X}^\top\right) = \sum_{i=1}^{n} \|\mathbf{x}_i\|^2$. So we get that the total distortion is

$$n \sum_{i=1}^{d} \lambda_i - n \sum_{i=1}^{r} \lambda_i = n \sum_{i=r+1}^{d} \lambda_i$$

So we have shown:

Proposition 6.4.2

The total distortion by PCA is

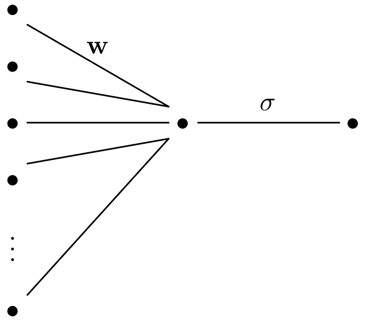
$$n \sum_{i=r+1}^{d} \lambda_i$$

So we want to choose an r small enough so that we can more easily deal with the dimension, but large enough so that the remainder $\sum_{i=r+1}^{d} \lambda_i$ is small.

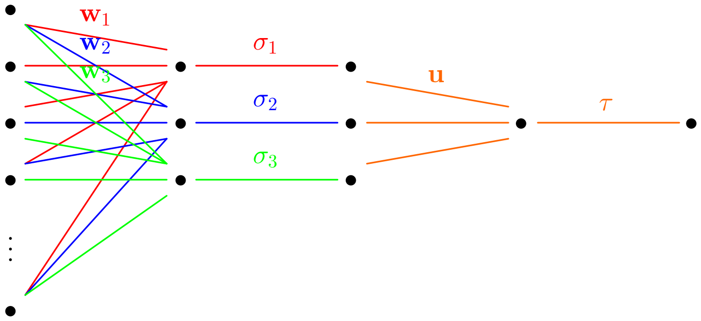
7 Deep Learning

Consider the following problem: we are given a stream of n bits, and we have to return whether the number of 1s is even or odd. For $n = 2$ this just reduces to exclusive-or (i.e. the answer is $x_1 \oplus x_2$). This is simple enough, but since the input is not linearly separable, it cannot be learned by the perceptron algorithm.

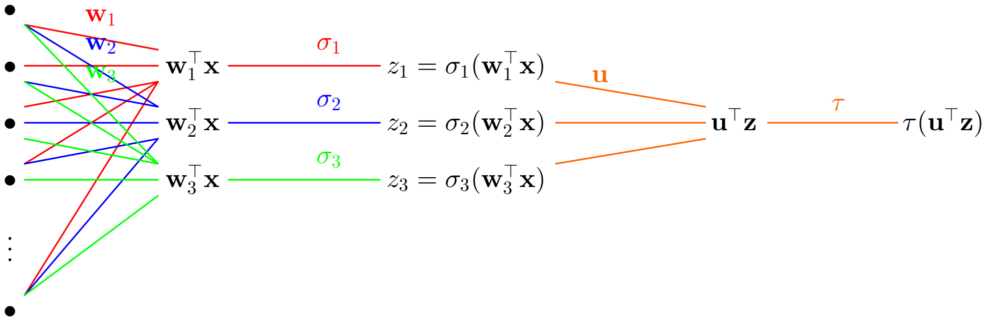
The idea of a perceptron is simple enough, recall that all we do is take all our features (a vector $\mathbf{x}$) and multiply by some weights (a vector $\mathbf{w}$) to get our output. We then alter the weight $\mathbf{w}$ in an attempt to minimize some loss function $\sum_{i=1}^n \ell(\mathbf{w}^\top \mathbf{x}_i, y_i)$. But as discussed, this requires the samples be linearly separable. The simplest way to remedy this is to add a step which applies some non-linear function σ. This can be described simply in the following image:



In this *perceptron unit*, we take an input vector $\mathbf{x}$, apply the weight $\mathbf{w}^\top \mathbf{x}$, then apply the nonlinearity $\sigma(\mathbf{w}^\top \mathbf{x})$. We can compose multiple perceptron units together, for example we can make a two-layer network:



If we want to keep track of the values at each node:


$\mathbf{w}_1$
$\mathbf{w}_2$
$\mathbf{w}_3$
$\mathbf{w}_1^\top \mathbf{x}$
$\mathbf{w}_2^\top \mathbf{x}$
$\mathbf{w}_3^\top \mathbf{x}$
σ_1
σ_2
σ_3
$z_1 = \sigma_1(\mathbf{w}_1^\top \mathbf{x})$
$z_2 = \sigma_2(\mathbf{w}_2^\top \mathbf{x})$
$z_3 = \sigma_3(\mathbf{w}_3^\top \mathbf{x})$
u
$\mathbf{u}^\top \mathbf{z}$
τ
$\tau(\mathbf{u}^\top \mathbf{z})$

Of course we don't want a single output, our network should have multiple output nodes, one for each class we are trying to classify by. In the final output we can apply a softmax nonlinearity to ensure that our outputs give valid probabilities. Some common choices for nonlinear activation functions (σ) are:

(1) Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$

(2) Tanh: $\tanh(x)$

(3) ReLU: $\max(0, x)$

(4) Leaky ReLU: $\max(0.1x, x)$

(5) Maxout: $\max(w_1 x + b_1, w_2 x + b_2)$

(6) ELU: $\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$

We also have choices for loss functions:

(**1**) 0-1 loss

(**2**) Hinge loss

(**3**) Cross-entropy loss

(**4**) Exponential loss