UMC203: AI and ML

Naman Mishra

January 2024

Contents

Lecture 01: Tue 09 Jan '24

0 The Course

Instructor: Prof. Chiranjib Bhattacharyya
Office: CSA, 254
Office hours: TBD
Lecture hours: TuTh 10:00–11:20

0.1 References

(i) *Pattern Classification* by Duda, Hart, and Stork

(ii) *Probabilistic Theory of Pattern Recognition* by Devroye, Györfi, and Lugosi

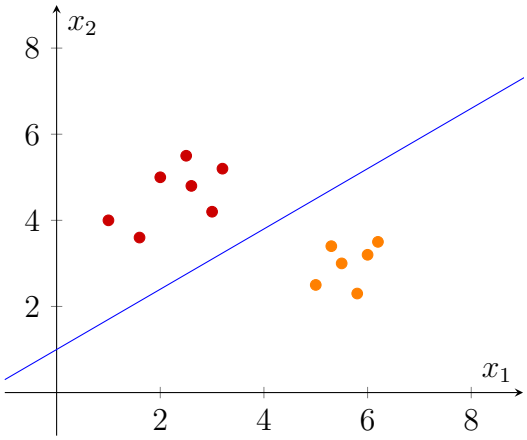
(iii) *Pattern Recognition and Machine Learning* by Bishop

1

Consider a machine which can measure the diameter of any fruit placed on it. Can the machine distinguish between an apple and an orange? Now suppose the machine also has the *capacity* to measure the weight of the fruit. Can it distinguish between an apple and an orange now?

$$\text{Fruit} \mapsto (x_1, x_2) \mapsto \{\text{Apple}, \text{Orange}\}$$

where x_1 is the diameter and x_2 is the weight. These are called *features*.



How do we measure how good a classifier is? This example has very few data points, so error is zero. Data is expensive, so accurate testing is expensive.

Let h be a classifier. We want to measure how good h is. We consider a random variable (of as yet unknown distribution) and compute the probability of error.

We consider this slightly more formally. Let the training data be

$$\mathcal{D} = \left\{(x^i, y^i)\right\}$$

where $x \in \mathbb{R}^2$ and $y \in \{1, -1\}$, where -1 and 1 represent apples and oranges respectively. Then h is a function from $\mathbb{R}^2$ to $\{-1, 1\}$. We wish to measure the probability $\Pr(h(X) \neq Y)$.

1.1 Probability Review

Suppose a coin is given with unknown probability of heads p. How do we estimate p? We flip the coin n times and count the number of heads n_H. Then we estimate p as $\hat{p} = \frac{n_H}{n}$.

The rationale behind this is the weak/strong law of large numbers.

Theorem 1.1 (Weak Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ. Then for any $\varepsilon > 0$,

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) \to 0$$

as $n \to \infty$.

Theorem 1.2 (Strong Law of Large Numbers)**.** Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ. Then

$$\Pr\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1$$

We have made several assumptions here. We know the structure of the problem at hand. For instance, we know that there is exactly one coin tossed each time. We have assumed that the coin tosses are independent. We have assumed that the probability of heads is the same for each toss.

Suppose we know the following in our earlier experiment:

- $\Pr(Y = 1)$
- $\Pr(Y = -1)$
- $\Pr(X = x \mid Y = 1)$
- $\Pr(X = x \mid Y = -1)$

Let $\eta(x) = \Pr(Y = 1 \mid X = x)$ given by Bayes' rule. Our rule for classification is

$$h(x) = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

This is called the *Bayes classifier*.

1.2 Multivariate Gaussians

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

where x is a d-dimensional column vector (so that the exponent is a scalar), μ is the mean, and Σ is the covariance matrix.

$$E[X] = \int_{x\in\mathbb{R}^d} x f(x)\,\mathrm{d}x$$
$$\Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$$

Reading assignment: Let A be a square symmetric real-valued matrix. What is known about the eigenvalues and positive definiteness of A?

Definition 1.3 (Positive definite)**.** A matrix $A_{n\times n}$ is *positive definite* if $u^\top A u > 0$ for all $u \in \mathbb{R}^n \setminus \{0\}$.

Exercise 1.4. Compute the Bayes classifier under the assumption that X under class 1 and class 2 has multivariate Gaussian distribution with means μ_1 and μ_2 with same covariance matrix Σ.

Lecture 02: Thu 11 Jan '24

We come back to apples and oranges.

$$\begin{aligned}\mathcal{X} &\subseteq \mathbb{R}^d \quad \text{instance space}\\ \mathcal{Y} &= \{-1, 1\} \quad \text{label space}\\ \mathcal{D} &= \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}\end{aligned}$$

We also know *priors*

$$\Pr(Y = 1) = p_1 \qquad \Pr(Y = -1) = 1 - p_1 =: p_2$$

and *class condition distributions*

$$\Pr(X = x \mid Y = 1) = f_1(x) \qquad \Pr(X = x \mid Y = -1) = f_2(x)$$

Remark. We will always write probabilities like this, but understand them to be densities whenever appropriate.

From Bayes' rule, we have the *posterior* $\eta\colon \mathcal{X} \to [0, 1]$ defined by

$$\begin{aligned}\eta(x) \coloneqq \Pr(Y = 1 \mid X = x) &= \frac{\Pr(X = x \mid Y = 1)\Pr(Y = 1)}{\Pr(X = x)}\\ &= \frac{f_1(x)p_1}{f_1(x)p_1 + f_2(x)p_2}\end{aligned}$$

We can then define the *Bayes classifier* as

$$\begin{aligned}h^*(x) &\coloneqq \operatorname{sgn}(2\eta(x) - 1)\\ &= \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2},\\ -1 & \text{otherwise}\end{cases}\\ &= \begin{cases} 1 & \text{if } f_1(x)p_1 > f_2(x)p_2,\\ -1 & \text{otherwise}\end{cases}\end{aligned}$$

For the specific case of multivariate Gaussians, *i.e.*, f_1 and f_2 of the form

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

with same covariance C but different means μ_1 and μ_2, we write

$$h^*(x) = \begin{cases} 1 & \text{if } \log \frac{\eta(x)}{1-\eta(x)} > 0 \\ -1 & \text{otherwise} \end{cases}$$

Now

$$\begin{aligned} \log \frac{\eta(x)}{1-\eta(x)} &= \log \frac{f_1(x)p_1}{f_2(x)p_2} \\ &= \log \frac{p_1}{p_2} - \frac{1}{2}\left((x-\mu_1)^\top C^{-1}(x-\mu_1)\right) \\ &\quad + \frac{1}{2}\left((x-\mu_2)^\top C^{-1}(x-\mu_2)\right) \\ &= \log \frac{p_1}{p_2} + (\mu_1-\mu_2)^\top C^{-1}x - \frac{1}{2}(\mu_1^\top C^{-1}\mu_1 - \mu_2^\top C^{-1}\mu_2) \\ &= w^\top x - b \end{aligned}$$

where $w = C^{-1}(\mu_1 - \mu_2)$ (since C is symmetric) and b is something. Thus $h^*(x) = \operatorname{sgn}(w^\top x - b)$.

Remark. $w^\top x = b$ is a hyperplane in $\mathbb{R}^d$, dividing the space into two half-spaces: $w^\top x < b$ and $w^\top x > b$. So a line is a very good guess for a classifier!

Exercise 1.5. Examine the special case of $C_1 = \sigma_1^2 I$ and $C_2 = \sigma_2^2 I$.

Solution. We have

$$\log \frac{f_1(x)}{f_2(x)} = d \log \frac{\sigma_2}{\sigma_1} - \frac{1}{2\sigma_1^2}\|x-\mu_1\|^2 + \frac{1}{2\sigma_2^2}\|x-\mu_2\|^2$$

Thus we choose class 1 when

$$d \log \sigma_1 + \frac{1}{2\sigma_1^2} < d \log \sigma_2 + \frac{1}{2\sigma_2^2}$$

and class 2 otherwise.

1.3 How Good is the Bayes Classifier?

We wish to compute the error $\Pr(h(X) \neq Y)$ for some rule h.

$$\begin{aligned} \Pr(h(X) \neq Y) &= E_{XY}\mathbf{1}_{h^*(X)\neq Y} \\ &= E_X E_{Y|X}\mathbf{1}_{h^*(X)\neq Y} \end{aligned}$$

but

$$E_{Y|X=x}\mathbf{1}_{h^*(X)\neq Y} = \begin{cases} 1-\eta(x) & \text{if } h(x)=1 \\ \eta(x) & \text{if } h(x)=-1 \end{cases} \qquad (*)$$
$$= \eta(x)\mathbf{1}_{h^*(x)=-1} + (1-\eta(x))\mathbf{1}_{h^*(x)=1}$$

It is clear from $(*)$ that whenever $\eta(x) > 1-\eta(x)$, setting $h(x)=1$ minimizes the error, and whenever $\eta(x) < 1-\eta(x)$, setting $h(x)=-1$ minimizes the error.

More rigorously, upon comparing with h^*,

$$\begin{aligned} E_{Y|X}(\mathbf{1}_{h(X)\neq Y} - \mathbf{1}_{h^*(X)\neq Y}) &= E_{Y|X}(\mathbf{1}_{h^*(X)=Y} - \mathbf{1}_{h^*(X)\neq Y}) \\ &= \eta(x)(\mathbf{1}_{h^*(X)=1} - \mathbf{1}_{h^*(X)=-1}) \\ &\quad + (1-\eta(x))(\mathbf{1}_{h^*(X)=-1} - \mathbf{1}_{h^*(X)=1}) \\ &= (2\eta(x)-1)(\mathbf{1}_{h^*(X)=1} - \mathbf{1}_{h^*(X)=-1}) \\ &= (2\eta(x)-1)(2\mathbf{1}_{h^*(X)=1} - 1) \end{aligned}$$

The second term is 1 when the first term is positive, and -1 when it is negative.

Thus

$$\begin{aligned} E_{XY}(\mathbf{1}_{h(X)\neq Y} - \mathbf{1}_{h^*(X)\neq Y}) &= E_X E_{Y|X}(\mathbf{1}_{h(X)\neq Y} - \mathbf{1}_{h^*(X)\neq Y}) \\ &= E_X|2\eta(x)-1| \geq 0. \end{aligned}$$

This proves that the Bayes classifier is the classifier with the lowest probability of error.
(This is theorem 2.1 in DGL.)

Lecture 03: Tue 16 Jan '24

1.4 Bayes' Decision Theory

We have a $x \in \mathbb{R}^d$ with label $y \in \{-1,1\}$. We predict $\hat{y} \in \{-1,1\}$. We have a loss function $\ell\colon \{-1,1\} \times \{-1,1\} \to \mathbb{R}_{\geq 0}$. We wish to minimize the expected loss

$$R(h) = E_{XY}\ell(h(X),Y).$$

This is minimized by

$$\tilde{h}(x) = \operatorname*{argmin}_h E_{Y|X=x}\ell(h(x),Y).$$

For a given instance, this rule chooses the label which yields the minimum loss.

Now

$$\begin{aligned} E_{Y|X=x}\ell(1,Y) &= \ell(1,1)\Pr(Y=1 \mid X=x) + \ell(1,-1)\Pr(Y=-1 \mid X=x) \\ &= \ell(1,1)\eta(x) + \ell(1,-1)(1-\eta(x)) \end{aligned}$$

Similarly

$$E_{Y|X=x}\ell(-1,Y) = \ell(-1,1)\eta(x) + \ell(-1,-1)(1-\eta(x))$$

$\tilde{h}$ minimises the loss if whenever $E_{Y|X=x}\ell(1,Y) < E_{Y|X=x}\ell(-1,Y)$, we choose $\tilde{h}(x) = 1$, and $\tilde{h}(x) = -1$ otherwise.

If $\ell(1,1) = \ell(-1,-1) = 0$ and $\ell(1,-1) = \ell(-1,1) = 1$, this reduces to the Bayes classifier.

1.5 Multi-Category Classification

Let us now extend the Bayes classifier to multiple classes. We have

$$\begin{aligned} &\mathcal{X} \subseteq \mathbb{R}^d \quad \text{instance space} \\ &\mathcal{Y} = [M] \quad \text{label space} \\ &\mathcal{D} = \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\} \end{aligned}$$

We also know priors

$$p_i = \Pr(Y = i)$$

and class condition distributions

$$f_i(x) = \Pr(X = x \mid Y = i)$$

We define the posteriors $\eta_i \colon \mathcal{X} \to [0,1]$ by

$$\eta_i(x) = \Pr(Y = i \mid X = x) = \frac{f_i(x)p_i}{\sum_{j=1}^{M} f_j(x)p_j}$$

and the Bayes classifier $h^* \colon \mathcal{X} \to \mathcal{Y}$ by

$$h^*(x) = \operatorname*{argmax}_{i \in \mathcal{Y}} \eta_i(x).$$

Suppose we also have a loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$. Then we have the *Bayes error rate*

$$E_{Y|X=x}\ell(\tilde{h}(x), Y) = \min_{i \in [M]} r_i(x)$$

where $r_i(x) = E_{Y|X=x}\ell(i, Y)$.

Lecture 04: Thu 18 Jan '24

We define the *Bayes risk*

$$R(h) = E_{XY}\ell(h(X), Y)$$

Theorem 1.6. Given the loss function $\ell = 1 - \delta$, where δ is the Kronecker delta, the Bayes classifier minimises the Bayes risk. That is, h

Proof. We have

$$\begin{aligned} r_i(x) &= \sum_{j=1}^{M} \ell(i,j)\eta_j(x) \\ &= \sum_{k\neq i} \eta_k(x) \\ &= 1 - \eta_i(x) \end{aligned}$$

We choose $\tilde{h}(x) = \operatorname{argmin}_i r_i(x)$. Then for all $j \neq i$,

$$\begin{aligned} r_i(x) &< r_j(x) \\ 1 - \eta_i(x) &< 1 - \eta_j(x) \\ \eta_j(x) &< \eta_i(x) \end{aligned}$$

For all $j \neq i$, $\tilde{h} = h^*$. □

For M category problem, define discriminant functions $g_i \colon \mathcal{X} \to \mathbb{R}$ for $i \in [M]$, and define the classifier $h(x) = \operatorname{argmax}_i g_i(x)$.

Let $f \colon [0,1] \to \mathbb{R}$ be any monotonically increasing function. Then $g_i = f \circ \eta_i$ works as a discriminant.

Suppose $\Pr(X = x \mid Y = i) = N(x \mid \mu_i, C_i)$. Then

$$\eta_i(x) = \frac{p_i N(x \mid \mu_i, C_i)}{P(x)}.$$

Then

$$\begin{aligned} \log \eta_i(x) &= \log p_i + \log N(x \mid \mu_i, C_i) - \log P(x) \\ &= \log p_i + \log \frac{1}{\left(\sqrt{2\pi}\right)^d |C|^{1/2}} - \frac{1}{2}(x-\mu_i)^\top C^{-1}(x-\mu_i) - \log P(x) \end{aligned}$$

If $C_i = C$ for all i, we drop the constants to get

$$g_i(x) = \log p_i - \frac{1}{2}(x - \mu_i)^\top C^{-1}(x - \mu_i).$$

Lecture 05: Tue 23 Jan '24

Exercise 1.7. Prove that

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ and}$$
$$\widehat{C} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$$

are the maximum likelihood estimators of μ and C for X_i i.i.d. from d-dimensional Gaussian distribution $N(\mu, C)$.

2 Fischer Discriminant

Suppose we know the mean and covariance of $X \mid Y = y$ for $y \in \{0, 1\}$. We wish to find w that maximizes

$$\frac{\left\|w^\top(\mu_0\mu_1)\right\|^2}{w^\top C_0 w + w^\top C_1 w}.$$

We can rewrite this as

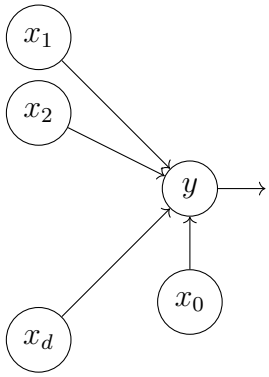
$$\frac{w^\top A w}{w^\top B w},$$

where $A = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^\top$ and $B = C_0 + C_1$. Note that both A and B are symmetric. Suppose that B is invertible and let L be a square root of B.

Lecture 06: Thu 25 Jan '24

3 Perceptron

We model a real biological neuron as an electrical circuit, so that it can be mimicked by silicon.



Let $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathcal{X}, y^{(i)} \in \{-1, 1\}, i \in [N]\}$.

Suppose there exists a w^* such that

$$\text{sgn}((w^*)^\top x^{(i)}) = y^{(i)} \text{ for all } i \in [N].$$

We know $y^{(i)}(w^{*\top} x^{(i)}) > 0$ for all $i \in [N]$.

Define the *margin* of w as

$$\gamma(w) = \min_{i \in [N]} \frac{y^{(i)}(w^\top x^{(i)})}{\|w\|}.$$

We want to maximize $\gamma(w)$.

We do this iteratively. Let $w^{(0)} = 0$.

Let $w^{(k)}$ be the current estimate of w^*. Let $(x^{(l)}, y^{(l)})$ be the first misclassified sample. Then, we update $w^{(k)}$ to $w^{(k+1)}$ by

$$w^{(k+1)} = w^{(k)} + y^{(l)} x^{(l)}.$$

If no such sample exists, then we are done.

$$\begin{aligned}
\|w_{k+1}\|^2 - \|w_k\|^2 &= (w^{(k+1)} - w^{(k)})^\top (w^{(k+1)} + w^{(k)}) \\
&= (y^{(l)} x^{(l)})^\top (2w^{(k)} + y^{(l)} x^{(l)}) \\
&= 2y^{(l)} (x^{(l)})^\top w^{(k)} + \left\|x^{(l)}\right\|^2 \\
&\leq \left\|x^{(l)}\right\|^2
\end{aligned}$$

and so

$$\left\|w^{(M)}\right\|^2 \leq MR^2. \tag{1}$$

On the other hand,

$$\begin{aligned}
(w^*)^\top (w^{(k+1)} - w^{(k)}) &= (w^*)^\top (y^{(l)} x^{(l)}) \\
&\geq \|w^*\| \gamma^*.
\end{aligned}$$

and so

$$(w^*)^\top w^{(M)} \geq M\|w^*\|\gamma^*$$

and from Cauchy-Schwarz,

$$\begin{aligned}\left\|w^*\right\|\left\|w^{(M)}\right\| &\geq M\|w^*\|\gamma^* \\ \left\|w^{(M)}\right\| &\geq M\gamma^*. \end{aligned} \tag{2}$$

Combining (1) and (2), we get

$$M^2\gamma^{*2} \leq MR^2 \iff M \leq \frac{R^2}{\gamma^{*2}}.$$

Thus the algorithm terminates in at most $\frac{R^2}{\gamma^{*2}}$ iterations.

3.1 Analysis

Let $\mathcal{D}$ be a training set of size N drawn i.i.d. from P. We denote this as $\mathcal{D} \sim P^N$.

Lecture 07: Thu 01 Feb '24

Setting:

$$\begin{aligned}&\mathcal{D} = \left\{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathcal{X}, y^{(i)} \in \mathcal{Y}, i \in [N]\right\} \\ &\mathcal{D} \sim P^{(N)} \\ &\mathcal{X} \subseteq \mathbb{R}^d \\ &\mathcal{Y} = \{-1, 1\}.\end{aligned}$$

Suppose there exists a $w^* \in \mathbb{R}^d$ such that for each $i \in [N]$, $\operatorname{sgn}(w^{*\top} x^{(i)}) = y^{(i)}$.
Then the algorithm described in the previous lecture will find such a w^* in at most $\frac{R^2}{\gamma^{*2}}$ iterations, where R is the maximum norm of $x^{(i)}$s and $\gamma^* = \min_{i\in[N]} \frac{|w^{*\top} x^{(i)}|}{\|w^*\|}$.

Let $\mathcal{D}$ be linearly separable and let the Perceptron algorithm return a classifier $h_{\mathcal{D}}^{(p)}$. We have risk $R(h_{\mathcal{D}}^{(p)}) = \Pr(h_{\mathcal{D}}^{(p)}(x) \neq y)$ (under what distribution?). We compute the *expected generalization error* by a classifier returned by the Perceptron algorithm acting on a linearly separable sample of size N drawn iid from P. That is, we compute

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim P^N}[R(h_{\mathcal{D}}^{(p)})].$$

This is hard!

We will instead compute the proxy $\overline{R}_{\mathcal{D}}^{LOO}(A)$, where A is an algorithm

acting on a sample $\mathcal{D}$ of size m, returning a classifier $h^A_{\mathcal{D}}$.
$\overline{R}^{LOO}_{\mathcal{D}}(A)$ is the *leave-one-out* error of A on $\mathcal{D}$. That is,

$$\overline{R}^{\mathrm{LOO}}_{\mathcal{D}}(A) = \frac{1}{m}\sum_{i=1}^{m}\left[h^A_{\mathcal{D}_{(i)}}\big(x^{(i)}\big) \neq y^{(i)}\right],$$

where $\mathcal{D}_{(i)} = \mathcal{D} \setminus \big\{(x^{(i)}, y^{(i)})\big\}$.

We want to compute the expected value of $\overline{R}^{LOO}_{\mathcal{D}}(A)$ over all samples $\mathcal{D}$ of size m drawn iid from P.

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim P^m}[\overline{R}^{\mathrm{LOO}}_{\mathcal{D}}(A)] = \frac{1}{m}\mathop{\mathbb{E}}_{\mathcal{D}\sim P^m}\sum_{i=1}^{m}\left[h^A_{\mathcal{D}_{(i)}}\big(x^{(i)}\big) \neq y^{(i)}\right]$$

Since the samples are iid, we have

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim P^m} = \mathop{\mathbb{E}}_{\mathcal{D}\sim P^{m-1}}\mathop{\mathbb{E}}_{(x^{(i)},y^{(i)})\sim P}$$

We first compute the inner expectation.

$$\begin{aligned}\mathop{\mathbb{E}}_{(x^{(i)},y^{(i)})\sim P}\left[h^A_{\mathcal{D}_{(i)}}\big(x^{(i)}\big) \neq y^{(i)}\right] &= \Pr\left(h^A_{\mathcal{D}_{(i)}}\big(x^{(i)}\big) \neq y^{(i)}\right)\\ &= R(h^A_{\mathcal{D}_{(i)}})\end{aligned}$$

So we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathcal{D}\sim P^m}[\overline{R}^{\mathrm{LOO}}_{\mathcal{D}}(A)] &= \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{\mathcal{D}_{(i)}\sim P^{m-1}} R(h^A_{\mathcal{D}_{(i)}})\\ &= \mathop{\mathbb{E}}_{\mathcal{D}\sim P^{m-1}} R(h^A_{\mathcal{D}}).\end{aligned}$$

Lecture 08: Tue 06 Feb '24

4 Convex Optimisation

Definition 4.1 (Convex function)**.** A set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0,1]$,

$$(1-\lambda)x + \lambda y \in C.$$

A function $f\colon C \to \mathbb{R}$ over a convex set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0,1]$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

Theorem 4.2. Let $f \in C^1(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Then f is convex iff

$$f(y) \geq f(x) + \langle \nabla f(x),\, y - x \rangle$$

Notation. Let A and B be symmetric matrices. We write $A \succeq B$ if $A - B$ is positive semidefinite.

Theorem 4.3. $\succeq$ is a partial order.

Proof.

- Reflexivity: $A - A = 0 \succeq 0$.
- Antisymmetry: $A - B \succeq 0$ and $B - A \succeq 0$ implies $A - B = 0$, since λ and $-\lambda$ are both nonnegative for each eigenvalue λ of the difference.
- Transitivity: Suppose $A \succeq B \succeq C$. Then for all u,

$$\begin{aligned} & \langle u,\, (A-B)u \rangle \geq 0 \\ & \langle u,\, (B-C)u \rangle \geq 0 \\ \implies \langle u,\, (A-C)u \rangle &= \langle u,\, (A-B+B-C)u \rangle \\ &= \langle u,\, (A-B)u \rangle + \langle u,\, (B-C)u \rangle \\ &\geq 0. \end{aligned}$$

□

Theorem 4.4. Let $f \in C^2(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Let $H(x) = (\operatorname{Hess} f)(x)$. Then f is convex iff

$$H(x) \succeq 0 \quad \forall\, x \in C.$$

Definition 4.5 (Convex optimisation problem)**.** Let $f \colon \mathbb{R}^d \to \mathbb{R}$ and $f_i \colon \mathbb{R}^d \to \mathbb{R}$ be convex functions for each $1 \leq i \leq m$. Let $(a_j)_{j=1}^n \subseteq \mathbb{R}^d$ and $(b_j)_{j=1}^n \subseteq \mathbb{R}$. The *convex optimisation problem* is to find

$$\min_{x \in \mathbb{R}^d} f(x) \quad \text{such that} \quad \begin{cases} f_i(x) \leq 0 \text{ for all } i \in [m], \\ \langle a_j,\, x \rangle = b_j \text{ for all } j \in [n]. \end{cases}$$

Definition 4.6 (Lagrangian)**.** Let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$. The *Lagrangian* of the convex optimisation problem is

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \mu_j \left(\langle a_j,\, x \rangle - b_j \right).$$

We say that x^* is a *KKT point* if there exist λ and μ such that

$$\begin{aligned} \nabla_x L(x^*, \lambda, \mu) &= 0, \\ \langle a_j,\, x^* \rangle - b_j &\leq 0 \quad \forall\, j \in [n], \\ f_i(x^*) &\leq 0 \quad \forall\, i \in [m], \\ \lambda_i f_i(x^*) &= 0 \quad \forall\, i \in [m]. \end{aligned}$$

Theorem 4.7. If x^* is a KKT point for the convex optimisation problem, then x^* is a minimiser of the problem (for most problems).

Example. Consider the convex optimisation problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{2} \|x - z\|^2 \quad \text{such that} \quad \langle w,\, x \rangle + b = 0.$$

The Lagrangian is

$$L(x, \mu) = \frac{1}{2} \|x - z\|^2 + \mu(\langle w,\, x \rangle + b).$$

The KKT conditions are

$$\begin{aligned} \nabla_x L(x^*, \mu) = x - z + \mu w &= 0, \\ \implies x^* &= z - \mu w, \\ \langle w,\, x^* \rangle + b &= 0 \\ \implies \langle w,\, z - \mu w \rangle + b &= 0 \\ \implies \langle w,\, z \rangle - \mu \|w\|^2 + b &= 0 \end{aligned}$$

So the minimiser is

$$x^* = z - \frac{(\langle w,\, z \rangle + b)}{\|w\|^2} w.$$

This is the orthogonal projection of z onto the hyperplane.

Lecture 09: Thu 08 Feb '24