UMC 203: AI and ML

Naman Mishra

January 2024

Contents

Lectures

The Course

Lecture 1.
Tuesday
January 09

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

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

(iv) *Foundations of Machine Learning* by Mohri, Rostamizadeh, and Talwalkar

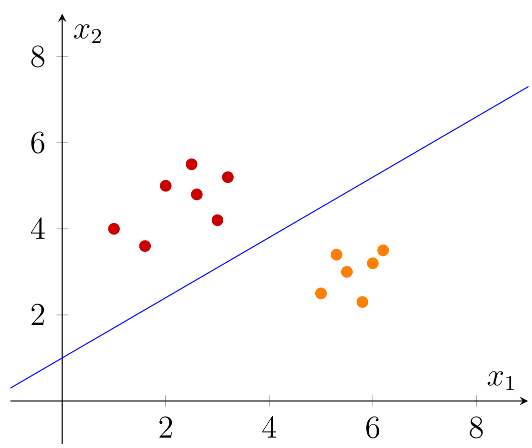
Chapter I

Bayes' Classifier

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 $\mathrm{P}(h(X) \neq Y)$.

I.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 $\hat{p} = n_H/n$.

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

Fact I.1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ. Then for any $\varepsilon > 0$,*

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

as $n \to \infty$.

Fact I.2 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ. Then*

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

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:

- $\mathrm{P}(Y = 1)$
- $\mathrm{P}(Y = -1)$
- $\mathrm{P}(X = x \mid Y = 1)$
- $\mathrm{P}(X = x \mid Y = -1)$

Let $\eta(x) = \mathrm{P}(Y = 1 \mid X = x)$ 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.*

Lecture 1: Classification and Bayes classifier

I.2 Multivariate Gaussians

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

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 I.3 (Definite matrix)**.** A matrix $A_{n\times n}$ is said to be *positive definite* if $u^\top Au > 0$ for all $u \in \mathbb{R}^n \setminus \{0\}$.

A is said to be *positive semidefinite* if $u^\top Au \geq 0$ for all $u \in \mathbb{R}^n$.

A is said to be *negative definite* and *negative semidefinite* if $-A$ is positive definite and positive semidefinite respectively.

Exercise I.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 2. Thursday January 11

We come back to apples and oranges.

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

We also know *priors*

$$\mathrm{P}(Y = 1) = p_1 \qquad \mathrm{P}(Y = -1) = 1 - p_1 =: p_2$$

and *class conditioned distributions*

$$\mathrm{P}(X = x \mid Y = 1) = f_1(x) \qquad \mathrm{P}(X = x \mid Y = -1) = f_2(x)$$

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

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

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

$$N(x \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\Big(-\frac{1}{2}\|x - \mu\|^2_{\Sigma^{-1}}\Big)$$

with same covariance Σ 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

$$\begin{aligned}\log \frac{\eta(x)}{1 - \eta(x)} &= \log \frac{f_1(x)p_1}{f_2(x)p_2}\\ &= \log \frac{p_1}{p_2} - \frac{1}{2}\langle x - \mu_1 \,|\, \Sigma^{-1} \,|\, x - \mu_1\rangle + \frac{1}{2}\langle x - \mu_2 \,|\, \Sigma^{-1} \,|\, x - \mu_2\rangle\\ &= \log \frac{p_1}{p_2} + \langle \mu_1 - \mu_2 \,|\, \Sigma^{-1} \,|\, x\rangle - \frac{1}{2}\Big(\|\mu_1\|^2_{\Sigma^{-1}} - \|\mu_2\|^2_{\Sigma^{-1}}\Big)\\ &= \langle w,\, x\rangle - b\end{aligned}$$

where $w = \Sigma^{-1}(\mu_1 - \mu_2)$ (since Σ is symmetric) and b is something. Thus $h^*(x) = \operatorname{sgn}(\langle w,\, x\rangle - b)$.

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

Lecture 2: Bayes classifier, multivariate gaussians and optimality

Exercise I.5. *Examine the special case of* $\Sigma_1 = \sigma_1^2 I$ *and* $\Sigma_2 = \sigma_2^2 I$.

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. ∎

I.3 How Good is the Bayes Classifier?

We wish to compute the error $\mathrm{P}(h(X) \neq Y)$ for some rule h.

$$\begin{aligned} \mathrm{P}(h(X) \neq Y) &= \mathop{\mathbf{E}}_{XY} \mathbf{1}_{h^*(X)\neq Y} \\ &= \mathop{\mathbf{E}}_{X} \mathop{\mathbf{E}}_{Y|X} \mathbf{1}_{h^*(X)\neq Y} \end{aligned}$$

but

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

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^*,

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

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

Thus

$$\underset{XY}{\mathbf{E}}(\mathbf{1}_{h(X)\neq Y} - \mathbf{1}_{h^*(X)\neq Y}) = \underset{X}{\mathbf{E}}\underset{Y|X}{\mathbf{E}}(\mathbf{1}_{h(X)\neq Y} - \mathbf{1}_{h^*(X)\neq Y})$$
$$= \underset{X}{\mathbf{E}}|2\eta(x) - 1| \geq 0.$$

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

I.4 Bayes' Decision Theory

Lecture 3. Tuesday January 16

We have an $x \in \mathbb{R}^d$ 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

$$R(h) = \underset{XY}{\mathbf{E}}\, \ell(h(X), Y).$$

This is minimized by

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

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

Now

$$\underset{Y|X=x}{\mathbf{E}} \ell(1, Y) = \ell(1,1)\,\mathrm{P}(Y = 1 \mid X = x) + \ell(1,-1)\,\mathrm{P}(Y = -1 \mid X = x)$$
$$= \ell(1,1)\eta(x) + \ell(1,-1)(1 - \eta(x))$$

Similarly

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

$\tilde{h}$ minimises the loss if whenever $\mathbf{E}_{Y|X=x}\, \ell(1, Y) < \mathbf{E}_{Y|X=x}\, \ell(-1, Y)$, 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.

I.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

$$p_i = \mathrm{P}(Y = i)$$

and class condition distributions

$$f_i(x) = \mathrm{P}(X = x \mid Y = i)$$

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

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

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

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

Lecture 4. Thursday January 18

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*

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

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

We define the *risk* of a classifier as

$$R(h) = \mathop{\mathbf{E}}_{XY} \ell(h(X), Y)$$

Theorem I.6. *Given the loss function $\ell = 1 - \delta$, where δ is the Kronecker delta, the Bayes classifier minimises the risk. That is,*

$$\tilde{h} = \operatorname*{argmin}_{h \colon \mathcal{X}\to\mathcal{Y}} R(h)$$

Lecture 4: Bayes error rate (continued), discriminant functions

Proof. Fix an $x \in \mathcal{X}$. The optimal classifier h^* has

$$\begin{aligned} h^*(x) &= \operatorname*{argmin}_{h\colon \mathcal{X}\to\mathcal{Y}} \mathop{\mathbf{E}}_{Y|X=x} \ell(h(x), Y) \\ &= \operatorname*{argmin}_h \sum_{i=1}^M \mathrm{P}(Y = i \mid X = x)\ell(h(x), i) \\ &= \operatorname*{argmin}_h \sum_{i\neq h(x)} \mathrm{P}(Y = k \mid X = x) \\ &= \operatorname*{argmin}_h \Big(1 - \eta_{h(x)}(x)\Big). \end{aligned}$$

Similarly we have

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

and we choose $\tilde{h}(x) = \operatorname{argmin}_i r_i(x)$.

Let $i = \tilde{h}(x)$. Then for all $j \neq i$,

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

Thus $1 - \eta_{h(x)}(x)$ is minimised by $h(x) = i$. Thus $h^*(x) = \tilde{h}(x)$. □

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 to give the Bayes classifier.

Suppose the class conditioned probabilities are given by $\mathrm{P}(X = x \mid Y = i) = N(x \mid \mu_i, \Sigma_i)$. Then

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

Then

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

Lecture 4: Bayes error rate (continued), discriminant functions

If $\Sigma_i = \Sigma$ for all i, we drop the constants to get the discriminant

$$g_i(x) = \log p_i - \frac{1}{2}\|x - \mu_i\|^2_{\Sigma^{-1}}.$$

Lecture 5. Tuesday January 23

Exercise I.7. *Prove that*

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \ \textit{and}$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})(X_i - \widehat{\mu})^\top$$

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

Chapter II

Fisher Discriminant

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

$$\frac{\langle w,\, \mu_0 - \mu_1 \rangle^2}{\langle w \,|\, \Sigma_0 \,|\, w \rangle + \langle w \,|\, \Sigma_1 \,|\, w \rangle},$$

in order to find a w along which the class means are well-separated with low variance. We can rewrite this as

$$\frac{\langle w \,|\, A \,|\, w \rangle}{\langle w \,|\, B \,|\, w \rangle},$$

where $A = |\mu_0 - \mu_1\rangle\langle \mu_0 - \mu_1| = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^\top$ and $B = \Sigma_0 + \Sigma_1$.

Definition II.1 (Rayleigh quotient)**.** The *Rayleigh quotient* of a Hermitian matrix M and a non-zero vector x is defined as

$$R(M; x) = \frac{\|x\|_M^2}{\|x\|^2}.$$

Theorem II.2. *The Rayleigh quotient is maximized at the largest eigenvalue of M, by the corresponding eigenvector.*

Proof. Let M be a Hermitian matrix and let $v_1, \ldots, v_n$ be an orthonormal eigenbasis of M with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. For any vector x, we can write

$$x = \sum_{i=1}^{n} \langle x,\, v_i \rangle v_i.$$

Then we have

$$\|x\|^2 = \sum_{i=1}^{n} |\langle x,\, v_i \rangle|^2,$$

and

$$\|x\|_M^2 = \sum_{i=1}^{n} |\langle x,\, v_i \rangle|^2 \lambda_i,$$

so that

$$R(M; x) = \frac{\sum_{i=1}^{n} |\langle x,\, v_i \rangle|^2 \lambda_i}{\sum_{i=1}^{n} |\langle x,\, v_i \rangle|^2}$$

which is a weighted average of the eigenvalues. Then clearly the maximum is achieved at the largest eigenvalue λ^* when x is a multiple of the corresponding eigenvector. □

Definition II.3 (Generalized Rayleigh quotient)**.** The *generalized Rayleigh quotient* of matrices A and B with a non-zero vector x is defined as

$$R(A, B; x) = \frac{\langle x \mid A \mid x \rangle}{\langle x \mid B \mid x \rangle},$$

where A and B are Hermitian and B is invertible.

Lemma II.4 (Square roots)**.** *Let B be a positive definite matrix. Then there exists a positive definite matrix L such that $L^2 = B$.*

Proof. Let $B = QDQ^\top$ be the eigenvalue decomposition of B. Then we can take $L = Q\sqrt{D}Q^\top$. □

Lemma II.5. *The generalized Rayleigh quotient $R(A, B; x)$ is equal to the Rayleigh quotient $R(L^{-1}AL^{-1}; Lx)$, where L is a square root of B.*

Proof.

$$\begin{aligned} R(A, B; x) &= \frac{\langle x \mid A \mid x \rangle}{\langle x \mid B \mid x \rangle} \\ &= \frac{\langle Lx \mid L^{-1}AL^{-1} \mid Lx \rangle}{\langle Lx,\, Lx \rangle} \\ &= R(L^{-1}AL^{-1}; Lx). \end{aligned}$$

□

Lecture 5: Bayes error rate (continued), discriminant functions

Theorem II.6. *The generalized Rayleigh quotient $R(A, B; x)$ is maximized at the largest eigenvalue of $B^{-1}A$, by the corresponding eigenvector.*

Proof. By lemma II.4, we can find a square root L of B. Then by lemma II.5, we have

$$R(A, B; x) = R(L^{-1}AL^{-1}; Lx),$$

so that by theorem II.2, the maximum is achieved at the largest eigenvalue of $L^{-1}AL^{-1}$ by Lx being the corresponding eigenvector.

But if Lx is an eigenvector of $L^{-1}AL^{-1}$ with eigenvalue λ, then $L^{-1}AL^{-1}Lx = L^{-1}Ax = \lambda Lx$, or $L^{-2}Ax = \lambda x$. Thus x is an eigenvector of $B^{-1}A$ with the same eigenvalue. □

Note that both A and B are symmetric. Suppose that B is invertible and let L be a square root of B.

Thus the above theorem gives us that the maximum of the Fisher criterion is achieved at the largest eigenvalue of $B^{-1}A$. Let this be λ^* and let w^* be the corresponding eigenvector. Then we have

$$\begin{aligned} &B^{-1}Aw^* = \lambda^* w^* \\ \implies\ &(\Sigma_0 + \Sigma_1)^{-1}|\mu_0 - \mu_1\rangle\langle\mu_0 - \mu_1|(w^*) = \lambda^* w^*. \end{aligned}$$

But $|v\rangle\langle w|(x) = \langle w,\, x\rangle v$ for any vector x. So

$$\langle\mu_0 - \mu_1,\, w^*\rangle(\Sigma_0 + \Sigma_1)^{-1}(\mu_0 - \mu_1) = \lambda^* w^*$$

This gives that w^* is a multiple of $(\Sigma_0+\Sigma_1)^{-1}(\mu_0-\mu_1)$. Taking $w^* = (\Sigma_0+\Sigma_1)^{-1}(\mu_0-\mu_1)$ gives the eigenvalue $\lambda^* = \langle\mu_0 - \mu_1 \,|\, (\Sigma_0 + \Sigma_1)^{-1} \,|\, \mu_0 - \mu_1\rangle$. Thus we have proven the following theorem.

Theorem II.7 (Fisher's criterion)**.** *The vector*

$$w^* = (\Sigma_0 + \Sigma_1)^{-1}(\mu_0 - \mu_1)$$

maximizes the Fisher criterion. The maximum value is

$$\lambda^* = \langle\mu_0 - \mu_1 \,|\, (\Sigma_0 + \Sigma_1)^{-1} \,|\, \mu_0 - \mu_1\rangle.$$

If B were not invertible, we would be solving the generalized eigenvalue problem $Bw = \lambda Aw$.

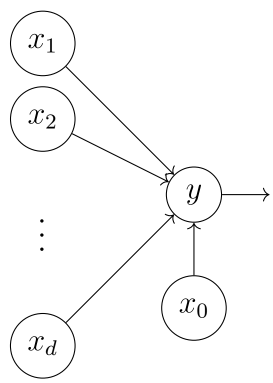
Lecture 5: Bayes error rate (continued), discriminant functions

Chapter III

Perceptron

Lecture 6. Thursday January 25

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



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

Definition III.1 (Margin)**.** Let $\mathcal{D}$ be as above. For any vector $w \in \mathbb{R}^d$, the *margin* of w with respect to $\mathcal{D}$ is

$$\gamma(w) = \min_{i \in [N]} \frac{y_i \langle w,\, x^{(i)} \rangle}{\|w\|}.$$

Suppose there exists a w^* such that

$$\operatorname{sgn}\langle w^*,\, x^{(i)} \rangle = y_i \text{ for all } i \in [N].$$

In other words, $\gamma(w^*) > 0$ or $y_i \langle w^*,\, x^{(i)} \rangle > 0$ for all $i \in [N]$.

We wish to find a w that maximizes $\gamma(w)$. For the time being, we'll be satisfied with any w that has positive margin.

III.1 The Algorithm

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

Let $w^{(k)}$ be the current weight vector. Let $(x^{(l)}, y_i)$ be the first misclassified sample. Then, we update $w^{(k)}$ to $w^{(k+1)}$ by

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

If no such sample exists, then we are done.

Since the numbering of samples is arbitrary, we can implement it as follows.

```
PERCEPTRON(𝒟):
  w ← 0
  for ever
      error ← ⊥
      for i ← 1 to N
          if y_i⟨w, x^(i)⟩ ≤ 0
              w ← w + y_i x^(i)
              error ← ⊤
      if ¬error
          return w
```

Notice that this does not break out of the current iteration when it finds a misclassified sample. It continues to check for more misclassified samples. For some reason, this gives *much* better results for the assignment problem.

III.2 Termination

Theorem III.2. *Let* $\mathcal{D} = \left\{(x^{(i)}, y_i) \mid x^{(i)} \in \mathcal{X}, y_i \in \{-1, 1\}, i \in [N]\right\}$ *be linearly separable by a weight vector* w^**. Then the Perceptron algorithm terminates in at most* $\frac{R^2}{\gamma^{*2}}$ *iterations, where*

$$R = \max_{i \in [N]} \|x^{(i)}\|, \text{ and}$$

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

Proof. Let $w^{(k)}$ misclassify $x^{(l)}$. Then

$$\begin{aligned}\|w_{k+1}\|^2 - \|w_k\|^2 &= \langle w_{k+1} - w_k,\, w_{k+1} + w_k\rangle \\ &= \langle y_l x^{(l)},\, w_{k+1} + w_k\rangle \\ &= y_l\langle x^{(l)},\, 2w_k + y_l x^{(l)}\rangle \\ &= 2y_l\langle x^{(l)},\, w^{(k)}\rangle + \|x^{(l)}\|^2 \\ &\le \|x^{(l)}\|^2\end{aligned}$$

since this sample is misclassified. Then for each iteration M, prior to which at least one sample is misclassified,

$$\|w^{(M)}\|^2 \le MR^2. \tag{1}$$

On the other hand,

$$\begin{aligned}\langle w^*,\, w^{(k+1)} - w^{(k)}\rangle &= \langle w^*,\, y_l x^{(l)}\rangle \\ &\ge \|w^*\|\gamma^*.\end{aligned}$$

and so

$$\langle w^*,\, w^{(M)}\rangle \ge M\|w^*\|\gamma^*.$$

From Cauchy-Schwarz,

$$\begin{aligned}\|w^*\|\,\|w^{(M)}\| &\ge M\|w^*\|\gamma^* \\ \|w^{(M)}\| &\ge M\gamma^*. \end{aligned} \tag{2}$$

Combining (1) and (2), we get

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

Thus no sample is misclassified after R^2/γ^{*2} iterations, so the algorithm terminates in at most this many iterations. □

III.3 Risk Analysis

Lecture 7. Thursday February 01

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

That is,

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

Suppose there exists a $w^* \in \mathbb{R}^d$ such that for each $i \in [N]$, $\operatorname{sgn}\langle w^*, x^{(i)}\rangle = y_i$. Then the algorithm described in the previous lecture will find a w that separates the samples in at most R^2/γ^{*2} iterations, where R is the maximum norm of $x^{(i)}$s (*radius*) and

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

Let $\mathcal{D}$ be linearly separable and let the Perceptron algorithm return a classifier $h_{\mathcal{D}}^{(p)}$. We have risk $R(h_{\mathcal{D}}^{(p)}) = \mathrm{P}_{X,Y\sim P}(h_{\mathcal{D}}^{(p)}(X) \neq Y)$ 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

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

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 $h_{\mathcal{D}}^{A}$.

III.3.1 Leave-One-Out Error

Definition III.3 (Leave-one-out error)**.** Let A be an algorithm that acts on a sample $\mathcal{D} = \left\{(x^{(i)}, y_i) \mid i \in [m]\right\}$ to return a classifier $h_{\mathcal{D}}^{A}$. The *leave-one-out error* of A on $\mathcal{D}$ is defined to be

$$\overline{R}_{\mathcal{D}}^{LOO}(A) := \frac{1}{m}\sum_{i=1}^{m}\left[h_{\mathcal{D}_{(i)}}^{A}\left(x^{(i)}\right) \neq y_i\right],$$

where $\mathcal{D}_{(i)} = \mathcal{D} \setminus \left\{(x^{(i)}, y_i)\right\}$.

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

$$\mathop{\mathbf{E}}_{\mathcal{D}\sim P^m}[\overline{R}_{\mathcal{D}}^{\mathrm{LOO}}(A)] = \frac{1}{m}\mathop{\mathbf{E}}_{\mathcal{D}\sim P^m}\sum_{i=1}^{m}\left[h_{\mathcal{D}_{(i)}}^{A}\left(x^{(i)}\right) \neq y_i\right]$$

Lecture 7: Generalization error of the perceptron algorithm

Since the samples are iid, we have

$$\underset{\mathcal{D}\sim P^m}{\mathbf{E}} = \underset{\mathcal{D}\sim P^{m-1}}{\mathbf{E}} \underset{(x^{(i)},y_i)\sim P}{\mathbf{E}}$$

We first compute the inner expectation.

$$\begin{aligned}\underset{(x^{(i)},y_i)\sim P}{\mathbf{E}}\Big[h^A_{\mathcal{D}_{(i)}}\Big(x^{(i)}\Big) \neq y_i\Big] &= \mathrm{P}\Big(h^A_{\mathcal{D}_{(i)}}\Big(x^{(i)}\Big) \neq y_i\Big) \\ &= R(h^A_{\mathcal{D}_{(i)}})\end{aligned}$$

So we have

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

Thus the expected leave-one-out error of an algorithm A acting on a sample of size m drawn iid from P is the expected risk of the classifier returned by A acting on a sample of size $m-1$ drawn iid from P. Thus we have proven the following theorem.

Theorem III.4. *Let A be an algorithm that acts on a sample $\mathcal{D}$ to return a classifier $h^A_{\mathcal{D}}$. Then the expected leave-one-out error of A acting on a sample of size m drawn iid from a probability distribution P is the expected risk of the classifier returned by A acting on a sample of size $m-1$ drawn iid from P.*

$$\underset{\mathcal{D}\sim P^m}{\boldsymbol{E}}[\overline{R}^{LOO}_{\mathcal{D}}(A)] = \underset{\mathcal{D}\sim P^{m-1}}{\boldsymbol{E}}[R(h^A_{\mathcal{D}})].$$

III.3.2 Perceptron's Leave-One-Out Error

Fix a sample set $\mathcal{D} \sim P^{N+1}$. Let i_k be the index of the sample misclassified in the kth iteration. Then the w returned by the Perceptron algorithm is

$$w^{(k)} = \sum_{j=1}^{M} y_{i_j} x^{(i_j)}$$

where $M = M(\mathcal{D})$ is the number of iterations. Let $I = \{i_1, i_2, \ldots, i_M\}$. Then for any index $i \notin I$, w correctly classifies $x^{(i)}$ at every iteration. Thus it makes no difference to the algorithm whether $x^{(i)}$ is in the sample or not.

The classifiers $h^{(p)}_{\mathcal{D}_{(i)}}$ and $h^{(p)}_{\mathcal{D}}$ are the same, and so the leave-one-out error

$$[h^{(p)}_{\mathcal{D}_{(i)}}(x^{(i)}) \neq y_i]$$

for this i is 0.

Thus the average leave-one-out error on $\mathcal{D}$ is at most

$$\frac{1}{N+1}\sum_{i\in U} 1 = \frac{M(\mathcal{D})}{N+1}.$$

By the previous bound on the number of iterations, we have

Theorem III.5. *The expected generalization error of the perceptron algorithm*

$$\underset{\mathcal{D}\sim P}{\boldsymbol{E}} R(h_{\mathcal{D}}^{(p)}) \leq \frac{M(D)}{N+1} \leq \frac{\rho(\mathcal{D})^2}{(N+1)\gamma^*(\mathcal{D})^2}$$

where $M(\mathcal{D})$ is the number of iterations that the perceptron algorithm takes to converge on $\mathcal{D}$, and $\rho(\mathcal{D})$ and $\gamma^*(\mathcal{D})$ are the radius and margin of $\mathcal{D}$ respectively.

Chapter IV

Convex Optimisation

Lecture 8. Tuesday February 06

Definition IV.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 $f: C \to \mathbb{R}$ over a convex set $C \subseteq \mathbb{R}^d$ is said to be *convex* if for all $x, y \in C$ and $\lambda \in [0,1]$,
$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

Fact IV.2. *Let $f \in C^1(C)$, where $C \subseteq \mathbb{R}^d$ is convex. Then f is convex iff*
$$f(y) - f(x) \geq \langle \nabla f(x),\, y - x \rangle$$
for all $x, y \in C$.

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

Proposition IV.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 if λ is an eigenvalue of $A - B$, then $-\lambda$ is an eigenvalue of $B - A$. But all eigenvalues of $A - B$ as well as $B - A$ are nonnegative, so $\lambda = 0$.

- Transitivity: Suppose $A \succeq B \succeq C$. Then for all u,

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

□

Fact IV.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 IV.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 $i \in [m]$. 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

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

IV.1 KKT Conditions

Definition IV.6. Let $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n$. The *Lagrangian* of the convex optimisation problem is

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

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

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

The first condition is the *stationarity* condition. The second and third conditions are the *primal feasibility* conditions. The final condition is the *complementary slackness* condition.

Fact IV.7. *If x^* is a KKT point for the convex optimisation problem, then x^* is a global minimiser.*

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 minimizer is

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

This is the orthogonal projection of z onto the hyperplane.

IV.2 Wolfe Dual

Definition IV.8 (Wolfe dual)**.** For a given convex optimisation problem P, the *Wolfe dual* problem D is

$$\max_{x, \lambda, \mu} L(x, \lambda, \mu) \quad \text{such that} \quad \begin{cases} \lambda \geq 0, \\ \nabla_x L(x, \lambda, \mu) = 0. \end{cases}$$

Theorem IV.9. *If x^* is a KKT point for the convex optimisation problem with Lagrange multipliers λ^* and μ^*, then (x^*, λ^*, μ^*) solves the Wolfe dual problem.*

Proof. First absorb the affine equality constraints into the convex inequality constraints. Suppose x^* is a KKT point with Lagrange multipliers λ^*. Note that (x^*, λ^*) is feasible for the Wolfe dual problem. Then

$$\begin{aligned} L(x^*, \lambda^*) &= f(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \\ &= f(x^*) \end{aligned}$$

by complementary slackness. Also note that by primal feasibility,

$$\begin{aligned} L(x^*, \lambda) &= f(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) \\ &\leq f(x^*) = L(x^*, \lambda^*). \end{aligned}$$

Let $f_0 = f$. Now since f_i, $i \in \{0, \ldots, m\}$, are convex, we have

$$f_i(x^*) \geq f_i(x) + \langle \nabla f_i(x),\, x^* - x \rangle$$

for all x.

Thus

$$\begin{aligned} L(x^*, \lambda) &= f(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) \\ &\geq f(x) + \langle \nabla f(x),\, x^* - x \rangle + \sum_{i=1}^m \lambda_i (f_i(x) + \langle \nabla f_i(x),\, x^* - x \rangle) \\ &= f(x) + \sum_{i=1}^m \lambda_i f_i(x) + \left\langle \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x),\, x^* - x \right\rangle \\ &= L(x, \lambda) + \langle x^* - x,\, \nabla_x L(x, \lambda) \rangle. \end{aligned}$$

Then if x is a feasible point for the Wolfe dual problem,

$$L(x^*, \lambda) \geq L(x, \lambda)$$

by the stationarity condition. Thus for all feasible x and λ,

$$L(x^*, \lambda^*) \geq L(x^*, \lambda) \geq L(x, \lambda). \qquad \square$$

Thus we can use the Wolfe dual to hunt for KKT points.

Chapter V

Large margin classification

Lecture 9. Thursday February 08

Let $\mathcal{D} = \left\{(x^{(i)}, y_i)\right\}_{i=1}^{m}$ be a linearly separable dataset. The perceptron algorithm finds a separating hyperplane, but there are many such hyperplanes. Which one is the best?

We can focus on the margin of the hyperplane. The margin is as defined in definition III.1. The hyperplane with the largest margin is deemed the best.

> **Definition V.1** (The SVM problem)**.** The *support vector machine* (SVM) problem is to find the hyperplane with the largest margin. That is, find w that solves
> $$\max_w \min_i \frac{y_i\langle w,\, x^{(i)}\rangle}{\|w\|}.$$

What about the more general classifiers using $\langle w,\, x\rangle + b$? We can append a constant 1 to each $x^{(i)}$ and append b to w. Hence we can restrict our attention to the case where $b = 0$.

Note that the objective function is homogeneous in w. So we can scale w such that $\min_i y_i\langle w,\, x^{(i)}\rangle = 1$. Then the problem becomes

$$\max_w \frac{1}{\|w\|} \quad \text{subject to} \quad \min_i y_i\langle w,\, x^{(i)}\rangle = 1.$$

When is $\min_i y_i\langle w,\, x^{(i)}\rangle = 1$? When $\langle w,\, y_i x^{(i)}\rangle \geq 1$ for all i, but also $\langle w,\, y_i x^{(i)}\rangle = 1$ for some i. What if $\langle w,\, y_i x^{(i)}\rangle > 1$ for all i? Then w can be shrunken to increase the objective while still satisfying the constraints. Thus the problem becomes

$$\max_w \frac{1}{\|w\|} \quad \text{subject to} \quad \langle w,\, y_i x^{(i)}\rangle \geq 1 \text{ for all } i.$$

But maximizing $1/\|w\|$ is the same as minimizing $\|w\|^2$. So we again rewrite the problem as

$$\min_w \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad \langle w,\, y_i x^{(i)}\rangle \geq 1 \text{ for all } i.$$

Note that $w \mapsto \|w\|^2$ is a strictly convex function. We have the Lagrangian

$$L(w, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^m \lambda_i(\langle w,\, y_i x^{(i)}\rangle - 1)$$

and so the KKT conditions

$$\nabla_w L(w, \lambda) = 0 \implies w = \sum_{i=1}^m \lambda_i y_i x^{(i)} \tag{V.1}$$

$$\langle w,\, y_i x^{(i)}\rangle \geq 1 \quad \text{for all } i, \tag{V.2}$$

$$\lambda_i(\langle w,\, y_i x^{(i)}\rangle - 1) = 0 \quad \text{for all } i. \tag{V.3}$$

If $\lambda_i > 0$, then $\langle w,\, y_i x^{(i)}\rangle = 1$. If $\langle w,\, y_i x^{(i)}\rangle > 1$, then $\lambda_i = 0$.

The $x^{(i)}$s for which $\lambda_i > 0$ are called the *support vectors*. These are at most the points for which $\langle w,\, y_i x^{(i)}\rangle = 1$.

Lecture 10. Friday February 09

Substituting equation (V.1) into the Lagrangian gives

$$\begin{aligned} L &= \frac{1}{2}\Big\|\sum_i \lambda_i y_i x^{(i)}\Big\|^2 - \sum_i \lambda_i \Big\langle \sum_j \lambda_j y_j x^{(j)},\, y_i x^{(i)}\Big\rangle + \sum_{i=1}^m \lambda_i \\ &= \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j \langle y_i x^{(i)},\, y_j x^{(j)}\rangle \end{aligned}$$

Thus using the Wolfe dual (section IV.2), the SVM problem is to solve

$$\max_\lambda \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i \lambda_j \Big\langle y_i x^{(i)},\, y_j x^{(j)}\Big\rangle \quad \text{subject to} \quad \lambda_i \geq 0$$

If we find such a λ, we have

$$w = \sum_{i=1}^m \lambda_i y_i x^{(i)}$$

and the classifier

$$h(x) = \operatorname{sgn}\langle w,\, x\rangle.$$

Except… this is **NOT** the SVM problem. The SVM problem does not absorb the constant b into the vector w.

Definition V.1 (The SVM problem)**.** The *support vector machine* (SVM) problem is to find the hyperplane with the largest margin. That is, find w and b that solve

$$\max_{w,b} \min_i \frac{y_i(\langle w,\, x^{(i)}\rangle + b)}{\|w\|}.$$

Notice that the norm in the denominator *does not* include b. Through much the same machinery as before, one arrives at the problem

$$\boxed{\max_\lambda \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i\lambda_j y_i y_j \langle x^{(i)},\, x^{(j)}\rangle \quad \text{subject to} \quad \begin{cases} \lambda_i \geq 0, \\ \sum_i \lambda_i y_i = 0. \end{cases}}$$

This determines w as before: $w = \sum_i \lambda_i y_i x^{(i)}$.

Note that the only dependence on $x^{(i)}$ is through the inner product. Thus we can use the *kernel trick* to solve the SVM problem in linearly non-separable cases.

Suppose $(x^{(i)}, y_i)$ are not linearly separable, but there is a transformation Φ such that $(\Phi(x^{(i)}), y_i)$ are linearly separable. Then we can apply SVM to the transformed dataset.

$$\max_\lambda \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i\lambda_j \Big\langle y_i\Phi(x^{(i)}),\, y_j\Phi(x^{(j)})\Big\rangle \quad \text{subject to} \quad \begin{cases} \lambda_i \geq 0, \\ \sum_i \lambda_i y_i = 0. \end{cases}$$

If we can compute $\Big\langle \Phi(x^{(i)}),\, \Phi(x^{(j)})\Big\rangle$, then we can solve the SVM problem for the transformed dataset.

V.1 Generalization Error

Lecture 12. Tuesday February 27

Theorem V.2 (FOML 5.4)**.** *Let $h_{\mathcal{D}}^{SVM}$ be the classifier returned by the SVM for a sample $\mathcal{D}$, and let $N_{SV}(\mathcal{D})$ be the number of support vectors that define $h_{\mathcal{D}}^{SVM}$. Then,*

$$\mathop{\boldsymbol{E}}_{\mathcal{D}\sim P^m}(R(h_{\mathcal{D}}^{SVM})) \leq \mathop{\boldsymbol{E}}_{\mathcal{D}\sim P^{m+1}}\left[\frac{N_{SV}(\mathcal{D})}{m+1}\right]$$

Proof. The proof is identical to that of theorem III.5, proceeding via the leave-one-out error. $\square$

If the training set error is zero, is the generalization error also zero?

Lecture 12: Linear SVM classifiers for linearly non-separable data: VC Dimension

Fact V.3. *Let $Z_1, \ldots, Z_n$ be iid random variables with* $\mathrm{P}(Z_i \in [a,b]) = 1$ *and* $\boldsymbol{E}[Z_i] = \mu$. *Then,*

$$\mathrm{P}\left(\left|\overline{Z} - \mu\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

where $\overline{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$.

We can use this to give probabilistic bounds on the generalization error using its expected value (since it is bounded between 0 and 1).

V.2 VC Dimension

Let $\mathcal{H}$ be a hypothesis class. That is, a set of functions from $\mathcal{X}$ to $\mathcal{Y}$. For our purposes, $\mathcal{Y} = \{-1, 1\}$.

Definition V.4 (Growth function)**.** The growth function $\Pi_{\mathcal{H}} \colon \mathbb{N} \to \mathbb{N}$ is defined by

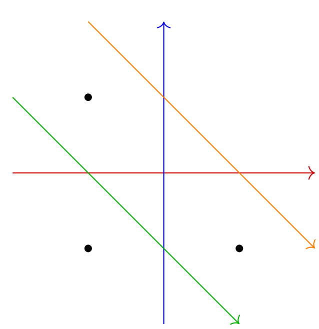
$$\Pi_{\mathcal{H}}(m) = \max_{x_1,\ldots,x_m \in \mathcal{X}} \#\{(h(x_1), \ldots, h(x_m)) \mid h \in \mathcal{H}\}$$

In other words, $\Pi_{\mathcal{H}}(m)$ is the maximum number of distinct ways in which m points can be classified by functions in $\mathcal{H}$.

Notation. We will denote the set of affine classifiers from $\mathbb{R}^d$ to $\{-1, 1\}$ by $\mathcal{L}_d$.

Example. If $\mathcal{H} = \mathcal{L}_2$, then

$$\Pi_{\mathcal{H}}(1) = 2$$
$$\Pi_{\mathcal{H}}(2) = 4$$
$$\Pi_{\mathcal{H}}(3) = 8$$



These four classifiers give four distinct ways to classify the given three points. Reversing these gives another four ways. There are only eight possible labelings, so $\Pi_{\mathcal{H}}(3) = 8$.

Lecture 12: Linear SVM classifiers for linearly non-separable data: VC Dimension

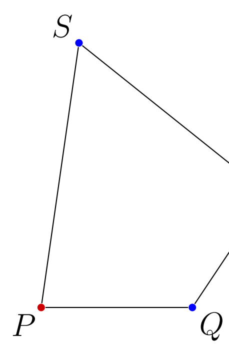
However, $\Pi_{\mathcal{H}}(4) < 16$. That is, no matter which 4 points we choose, we can't find 16 distinct classifications of them by functions in $\mathcal{H}$. In other words, for any 4 points, there exists a labeling of them that cannot be achieved by any function in $\mathcal{H}$.

Theorem V.5. $\Pi_{\mathcal{L}_2}(4) < 16$.

Proof. Let P, Q, R and S be any four points, colored red or blue. The key observation is that if a line L separates the red points from the blue points, then for any two points A and B, L intersects the line segment $\overline{AB}$ iff A and B are colored differently.

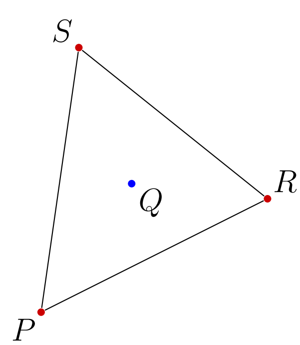
If any three points P, Q and R are collinear in that order, color them red, blue and red respectively. Then any line separating the red points from the blue points must pass through both $\overline{PQ}$ and $\overline{QR}$. The only line that does this is the line $\overline{PR}$ itself, which will assign the same color to each of these.

Now suppose that P, Q, R and S are such that no three are collinear. If they form a convex quadrilateral, color them alternately red and blue.



Any line separating the red points from the blue points must intersect every side of the quadrilateral, which is not possible.

If they form a non-convex quadrilateral, the convex hull must be a triangle. Color the points of the triangle red, and the interior point blue.



A separating plane can pass through none of the sides of the triangle, so it cannot enter the interior of the triangle at all, and thus cannot separate Q from the other points. $\square$

Definition V.6 (Shattering)**.** A hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is said to *shatter* a set $C \subseteq \mathcal{X}$ if for every labelling of C by $\mathcal{Y}$, there exists a function $h \in \mathcal{H}$ that achieves that labelling. That is,

$$\forall y \in \mathcal{Y}^C \; \exists h \in \mathcal{H} \; \forall x \in C (h(x) = y(x))$$

Example. From the above theorem, we conclude that $\mathcal{L}_2$ shatters no set of four points.

From the example preceding it, we conclude that the set of linear classifiers in $\mathbb{R}^2$ shatters that particular set of three points (and indeed, any set of three points that are not collinear).

Definition V.7 (VC-dimension)**.** The *VC-dimension* of a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the size of the largest set that can be shattered by $\mathcal{H}$. That is,

$$\mathrm{VC}(\mathcal{H}) = \max\{m \mid \Pi_{\mathcal{H}}(m) = |\mathcal{Y}|^m\}$$

Example. The VC-dimension of the set of linear classifiers in $\mathbb{R}^2$ is 3. This is because it shatters at least one set of three points, but no set of four points.

Theorem V.8. *The VC-dimension of the set of linear classifiers from $\mathbb{R}^d$ to $\{-1, 1\}$ is $d + 1$.*

Proof. Induction. For $d = 1$, the points -1 and 1 can obviously be shattered, using the affine maps $x \mapsto x$, $x \mapsto -x$, $x \mapsto x + 2$ and $x \mapsto -x - 2$.

Also, any three points are collinear, so they cannot be shattered by the same argument as in the proof of theorem V.5. Thus $\mathrm{VC}(\mathcal{L}_1) = 2$.

Suppose that $\mathrm{VC}(\mathcal{L}_{d-1}) = d$. Then let $P_1, \ldots, P_d$ be a set shattered by $\mathcal{L}_{d-1}$. Let $Q_i = (P_i, 0)$ for $i = 1, \ldots, d$, and $Q_{d+1} = (0, \ldots, 0, 1)$. We claim that the set $\{Q_1, \ldots, Q_{d+1}\}$ can be shattered by $\mathcal{L}_d$.

Fix a coloring y of $\{Q_1, \ldots, Q_{d+1}\}$. Consider the same coloring applied to $\{P_1, \ldots, P_d\}$ (each P_i colored the same as Q_i). Let $h(x) = \mathrm{sgn}(\langle w, x\rangle + b)$ be the classifier that achieves this coloring. WLOG assume that Q_{d+1} is colored $+1$. Let $w' = (w, 1-b)$. Then $h(x) = \mathrm{sgn}(\langle w', x\rangle + b)$ achieves the coloring y of $\{Q_1, \ldots, Q_{d+1}\}$. Thus $\mathrm{VC}(\mathcal{L}_d) \geq d + 1$.

Lecture 12: Linear SVM classifiers for linearly non-separable data: VC Dimension

To show that $\mathrm{VC}(\mathcal{L}_d) < d+2$, consider any set of $d+2$ points in $\mathbb{R}^d$. Suppose that they are shattered by $\mathcal{L}_d$. Fix a coloring y of these points. Consider the same coloring applied to any $d+1$ points, viewed as points in $\mathbb{R}^{d-1}$. Since there exists a classifier that achieves this coloring in $\mathbb{R}^d$, its restriction to $\mathbb{R}^{d-1}$ achieves the same coloring in $\mathbb{R}^{d-1}$. But this is impossible, since $\mathrm{VC}(\mathcal{L}_{d-1}) = d$. Thus $\mathrm{VC}(\mathcal{L}_d) < d+2$.

Winduction. □

Fact V.9. *Let*

$$\mathcal{X} = \left\{x \in \mathbb{R}^d \mid \|x\| \le R\right\}, \text{ and}$$
$$\mathcal{H}_B = \left\{h \in \{-1,1\}^{\mathcal{X}} \mid h = \operatorname{sgn}(\langle w, \cdot\rangle + b) \text{ for some } \|w\| \le B\right\}$$

be a class of linear classifiers on the ball $\mathcal{X}$. Then

$$\mathrm{VC}(\mathcal{H}_B) \le B^2R^2$$

Why are we interested in the VC-dimension at all?

Fact V.10. *Let $\mathcal{H}$ be a family of functions taking values in $\{-1,1\}$ with VC-dimension V. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in \mathcal{H}$.*

$$R(h) \le R_{emp}(h) + \sqrt{\frac{V}{N}(\log N - \log\delta)}.$$

What the fuck does this mean?

Corollary V.11. *For any $h \in \mathcal{H}_B$ defined in fact V.9, with probability at least $1-\delta$,*

$$R(h) \le R_{emp}(h) + O\left(\frac{RB}{\sqrt{N}}\right)$$

Where did the $\log N$ go?

This motivates the following formulation of the SVM problem for linearly non-separable data, since smaller w gives smaller bounds on the error.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(1 - y_i(\langle w, x^{(i)}\rangle + b)\right)_+$$

where $(x)_+ = x[x \ge 0] = 0 \vee x = \max(0,x)$, and C is a penalty for wrong answers.

V.3 Nonseparable SVM

Lecture 13. Tuesday February 27

We can rewrite this more conveniently (without the ugly max function) by introducing *slack variables* $\xi_i \geq 0$ for each $i \in [n]$.

$$\boxed{\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{subject to} \quad \begin{cases} y_i\Big(\langle w,\, x^{(i)}\rangle + b\Big) \geq 1 - \xi_i, \\ \xi_i \geq 0. \end{cases}}$$

Now **SOLVE!**

The Lagrangian is

$$\begin{aligned} L(w, b, \xi, \lambda^{(1)}, \lambda^{(2)}) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\quad - \sum_{i=1}^{n}\lambda_i^{(1)}\Big(y_i(\langle w,\, x^{(i)}\rangle + b) - 1 + \xi_i\Big) - \sum_{i=1}^{n}\lambda_i^{(2)}\xi_i. \end{aligned}$$

For the KKT point, the stationary conditions are

$$0 = \nabla_w L = w - \sum_{i=1}^{n}\lambda_i^{(1)} y_i x^{(i)}, \tag{V.4}$$

$$0 = \nabla_b L = -\sum_{i=1}^{n}\lambda_i^{(1)} y_i, \tag{V.5}$$

$$0 = \nabla_\xi L = C - \lambda^{(1)} - \lambda^{(2)}. \tag{V.6}$$

The complementary slackness conditions are

$$0 = \lambda_i^{(1)}\Big(y_i(\langle w,\, x^{(i)}\rangle + b) - 1 + \xi_i\Big), \tag{V.7}$$

$$0 = \lambda_i^{(2)}\xi_i. \tag{V.8}$$

V.3.1 Wolfe Dual

The Wolfe dual is

$$\max_{w,b,\xi,\lambda^{(1)},\lambda^{(2)}} L(w, b, \xi, \lambda^{(1)}, \lambda^{(2)}) \quad \text{subject to} \quad \begin{cases} \lambda^{(1)} \geq 0, \\ \lambda^{(2)} \geq 0, \\ \lambda_i^{(1)} + \lambda_i^{(2)} = C, \\ w = \sum_{i=1}^{n}\lambda_i^{(1)} y_i x^{(i)}, \\ \sum_{i=1}^{n}\lambda_i^{(1)} y_i = 0. \end{cases}$$

Lecture 13: Quadradic programming formulation of soft-margin SVM

Substituting equations (V.5) and (V.6) into L,

$$L^* = -\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i^{(1)} \langle w,\, y_i x^{(i)} \rangle + \sum_{i=1}^{n} \lambda_i^{(1)}.$$

Substituting equation (V.4),

$$L^* = \sum_i \lambda_i^{(1)} - \frac{1}{2} \sum_{i,j} \lambda_i^{(1)} \lambda_j^{(1)} y_i y_j \langle x^{(i)},\, x^{(j)} \rangle.$$

We have seen this before! In the linearly separable case, the only difference was that $\lambda_i^{(1)}$ were positive unrestricted, but here they are bounded above by C, because of equation (V.6). Thus the Wolfe dual boils down to

$$\boxed{\max_{0 \le \lambda^{(1)} \le C} \sum_i \lambda_i^{(1)} - \frac{1}{2} \sum_{i,j} \lambda_i^{(1)} \lambda_j^{(1)} y_i y_j \langle x^{(i)},\, x^{(j)} \rangle.}$$

Equation (V.6) is very interesting. For each i, $\lambda_i^{(1)} + \lambda_i^{(2)} = C$.

- If $\lambda_i^{(1)} = 0 \iff \lambda_i^{(2)} = C$, then $\xi_i = 0$ by equation (V.8). This gives $y_i(\langle w,\, x^{(i)} \rangle + b) \ge 1$ for the constraints to hold.
- If $0 < \lambda_i^{(1)} < C \iff 0 < \lambda_i^{(2)} < C$, then $\xi_i = 0$ by equation (V.8). But from equation (V.7), $y_i(\langle w,\, x^{(i)} \rangle + b) = 1$.
- If $\lambda_i^{(1)} = C \iff \lambda_i^{(2)} = 0$, then $0 \le \xi_i = (1 - y_i(\langle w,\, x^{(i)} \rangle + b)) \vee 0$.

Also note that $\xi_i > 0$ is possible only in the last case, $\lambda_i^{(1)} = C$ or $\lambda_i^{(2)} = 0$. This is also the only case where $y_i\Big(\langle w,\, x^{(i)} \rangle + b\Big) < 1$. This makes sense, because for the objective function to be minimized, ξ_i needs to be as small as possible. There is no need to have a positive ξ_i if the constraints are satisfied without it.

Chapter VI

Kernel Functions

Lecture 11.
Tuesday
February 27

Chapter VII

Regression

Lecture 14.
Monday
March 04

We move onto *continuous* data. Consider the data

$$\mathcal{D} = \left\{(x^{(i)}, y_i)\right\}_{i=1}^{n} \subseteq \mathcal{X} \times \mathcal{Y}$$
$$\mathcal{X} = \mathbb{R}^d$$
$$\mathcal{Y} = \mathbb{R} \quad \longleftarrow \text{woah, continuous!}$$

We again wish to find a classifier $f : \mathcal{X} \to \mathcal{Y}$ that minimizes some notion of loss. We will first focus on the affine case, that is, $f(x) = \langle w, x \rangle + b$. We can again employ the trick of appending a 1 to the input vector, so that we can focus on the linear case $f(x) = \langle w, x \rangle$.

VII.1 Least Squares Regression

The most natural choice of loss function for continuous data is the squared error loss, given by

$$\ell(\hat{y}, y) = (\hat{y} - y)^2$$

We can solve for the optimal w by minimizing the empirical risk, that is,

$$w_{LS} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \left\langle w, x^{(i)} \right\rangle\right)^2.$$

There are no constraints, so we need not worry about the KKT business.

SOLVE!

Define

$$D = \begin{bmatrix} x^{(1)} & \cdots & x^{(n)} \end{bmatrix}^\top \in \mathbb{R}^{n \times d} \qquad t = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

Then the empirical risk can be written as

$$\begin{aligned} R(w) &= \frac{1}{2} \begin{bmatrix} \langle x^{(1)},\, w \rangle - y_1 \\ \vdots \\ \langle x^{(n)},\, w \rangle - y_n \end{bmatrix}^2 \\ &= \frac{1}{2} \|Dw - t\|^2 \end{aligned}$$

Note that this is convex, since

$$\begin{aligned} \nabla R(w) &= D^\top (Dw - t) \\ \nabla^2 R(w) &= D^\top D \succeq 0 \end{aligned}$$

Thus the minimizer is given by

$$\boxed{w_{LS} = (D^\top D)^{-1} D^\top t}$$

What if $D^\top D$ is not invertible? More realistically, what if $\det\left(D^\top D\right)$ is very small? Adding a small multiple of the identity matrix to $D^\top D$ can help.

VII.2 Ridge Regression

Instead of minimizing the empirical risk, we can minimize the empirical risk plus a regularization term.

$$w_{RR} = \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{2} \sum_i \left(y_i - \left\langle w,\, x^{(i)} \right\rangle \right)^2 + \frac{\lambda}{2} \|w\|^2$$

where $\lambda > 0$ is the regularization parameter. This is called *Ridge Regression*.

We now have

$$\begin{aligned} \nabla R(w) &= \frac{1}{2} D^\top (Dw - t) + \lambda w \\ \nabla^2 R(w) &= D^\top D + \lambda I \end{aligned}$$

Lecture 14: Least squares and Ridge regression for linear models

This is still convex, and now the minimizer is

$$\boxed{w_{RR} = (D^\top D + \lambda I)^{-1} D^\top t}$$

VII.3 Optimal Classifier

Suppose that $\mathcal{D}$ is drawn from a distribution with

$$Y = f^*(X) + \varepsilon,$$

where ε is a random variable with mean 0 and variance σ^2. That is,

$$\mathrm{P}(Y \le y \mid X = x) = \mathrm{P}(f^*(x) + \varepsilon \le y).$$

If ε is Gaussian, then

$$\mathrm{P}(Y \mid X = x) = N(f^*(x), \sigma^2)$$

Let f be a classifier. Then the risk of f is given by

$$\begin{aligned} R(f) &= \mathop{\mathbf{E}}_{X,Y} \ell(f(X), Y) \\ &= \mathop{\mathbf{E}}_{X,Y} (f(X) - Y)^2 \\ &= \mathop{\mathbf{E}}_{X} \mathop{\mathbf{E}}_{Y|X} (Y - f(X))^2 \\ &= \mathop{\mathbf{E}}_{X} \mathop{\mathrm{Var}}_{Y|X} Y + \mathop{\mathbf{E}}_{X} (\mathop{\mathbf{E}}_{Y|X} Y - f(X))^2 \\ &\ge \mathop{\mathbf{E}}_{X} \mathop{\mathrm{Var}}_{Y|X} Y \end{aligned}$$

The equality holds iff $f(X) = \mathbf{E}_{Y|X} Y$ almost surely. Thus the optimal classifier is given by

$$f_B(x) = \mathop{\mathbf{E}}_{Y|X} Y$$

This computation is due to the following general result:

$$\begin{aligned} \mathbf{E}(X - a)^2 &= \mathbf{E}(X - \mathbf{E} X + \mathbf{E} X - a)^2 \\ &= \mathbf{E}(X - \mathbf{E} X)^2 + \mathbf{E}[2(X - \mathbf{E} X)(\mathbf{E} X - a)] + \mathbf{E}(\mathbf{E} X - a)^2 \\ &= \mathrm{Var}\, X + (\mathbf{E} X - a)^2 \end{aligned} \tag{VII.1}$$

This is called the *bias-variance decomposition*.

For $Y = f^*(X) + \varepsilon$, we have

$$\mathop{\mathbf{E}}_{X} \mathop{\mathrm{Var}}_{Y|X} Y = f^*(X) \qquad\qquad \mathop{\mathrm{Var}}_{Y|X} Y = \sigma^2$$

so that

$$R(f) = \sigma^2 + \mathop{\mathbf{E}}_{X}(f(X) - f^*(X))^2 \geq \sigma^2$$

The optimal classifier is, unsurprisingly, f^*.

VII.4 Generalization Errors

Lecture 15.
Tuesday
March 05

Let $f(x;\mathcal{D})$ be the classifier returned on the data set $\mathcal{D}$. The generalization error is given by

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}} R(f(x;\mathcal{D})) &= \mathop{\mathbf{E}}_{\mathcal{D}}\mathop{\mathbf{E}}_{X,Y} \ell(f(X;\mathcal{D}),Y) \\ &= \mathop{\mathbf{E}}_{X,Y}\mathop{\mathbf{E}}_{\mathcal{D}} \ell(f(X;\mathcal{D}),Y)\end{aligned}$$

From equation (VII.1) again,

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{D}} \ell(f(X;\mathcal{D}),Y) &= \mathop{\mathbf{E}}_{\mathcal{D}}(f(X;\mathcal{D}) - Y)^2 \\ &= \mathop{\mathrm{Var}}_{\mathcal{D}} f(X;\mathcal{D}) + (\mathop{\mathbf{E}}_{\mathcal{D}} f(X;\mathcal{D}) - Y)^2\end{aligned}$$

So the generalization error is given by

$$\begin{aligned}&\mathop{\mathbf{E}}_{X,Y}\mathop{\mathrm{Var}}_{\mathcal{D}|X} f(X;\mathcal{D}) + \mathop{\mathbf{E}}_{X,Y}(\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) - Y)^2 \\ &\qquad = \mathop{\mathbf{E}}_{X}\mathop{\mathrm{Var}}_{\mathcal{D}|X} f(X;\mathcal{D}) + \mathop{\mathbf{E}}_{X}\mathop{\mathbf{E}}_{Y|X}(Y - \mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}))^2\end{aligned}$$

and again using equation (VII.1),

$$\begin{aligned}&= \mathop{\mathbf{E}}_{X}\mathop{\mathrm{Var}}_{\mathcal{D}|X} f(X;\mathcal{D}) + \mathop{\mathbf{E}}_{X}\mathop{\mathrm{Var}}_{Y|X} Y + \mathop{\mathbf{E}}_{X}\left(\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) - \mathop{\mathbf{E}}_{Y|X} Y\right)^2 \\ &= \underbrace{\mathop{\mathbf{E}}_{X}\mathop{\mathrm{Var}}_{\mathcal{D}|X} f(X;\mathcal{D})}_{\text{variance}} + \underbrace{\mathop{\mathbf{E}}_{X}\mathop{\mathrm{Var}}_{Y|X} Y}_{\text{noise}} + \underbrace{\mathop{\mathbf{E}}_{X}\left(\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) - f_B(X)\right)^2}_{\text{bias}^2}\end{aligned}$$

For $Y = f^*(X) + \varepsilon$, this becomes

$$\mathop{\mathbf{E}}_{X}\mathop{\mathrm{Var}}_{\mathcal{D}|X} f(X;\mathcal{D}) + \sigma^2 + \mathop{\mathbf{E}}_{X}\left(\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) - f^*(X)\right)^2$$

Finally, we come to the linear case, with specific algorithms.

VII.4.1 Least Squares

We have

$$Y = \langle w,\, X\rangle + \varepsilon$$

and

$$f(x;\mathcal{D}) = \langle w_{LS},\, x\rangle$$

where

$$w_{LS} = (D^\top D)^{-1} D^\top t$$

We have three terms to compute:

- **Bias:**

$$\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) - \langle w,\, X\rangle$$

- **Variance:**

$$\mathop{\mathrm{Var}}_{\mathcal{D}|X} f(X;\mathcal{D})$$

- **Noise:** This we already know to be

$$\mathop{\mathbf{E}}_{X} \mathop{\mathrm{Var}}_{Y|X} Y = \sigma^2$$

Since $t = Dw + \varepsilon$, $\mathbf{E}[t] = Dw$. So

$$\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) = \mathop{\mathbf{E}}_{\mathcal{D}|X} \langle X,\, w_{LS}\rangle = \langle X,\, \mathop{\mathbf{E}}_{\mathcal{D}|X} w_{LS}\rangle = \langle X,\, \mathop{\mathbf{E}}_{\mathcal{D}|X} D^\top D^{-1} D^\top D w\rangle = \langle X,\, w\rangle.$$

Thus the bias is 0.

The variance is hard to compute. We will find an estimate using the *fixed design setting*. That is, we will assume that the data set $\mathcal{D}$ precisely represents the distribution of X.

$$\mathrm{P}(X = x) = \frac{1}{n}\sum_{i=1}^{n} [x = x^{(i)}].$$

Then

$$
\begin{aligned}
\operatorname*{Var}_{\mathcal{D}|X} f(X;\mathcal{D}) &= \mathop{\mathbf{E}}_{\mathcal{D}|X}\Big(f(X;\mathcal{D}) - \mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D})\Big)^2 \\
&= \mathop{\mathbf{E}}_{\mathcal{D}|X}\big(f(X;\mathcal{D}) - \langle w,\, X\rangle\big)^2 \\
&= \mathop{\mathbf{E}}_{\mathcal{D}|X}\langle w_{LS} - w,\, X\rangle^2 \\
&= \mathop{\mathbf{E}}_{\varepsilon}\frac{1}{n}\sum_{i=1}^{n}\langle (D^\top D)^{-1}D^\top\varepsilon,\, x^{(i)}\rangle^2 && \text{(fixed design)} \\
&= \frac{1}{n}\mathop{\mathbf{E}}_{\varepsilon}\varepsilon^\top D(D^\top D)^{-1}\sum_{i=1}^{n} x^{(i)}x^{(i)\top}(D^\top D)^{-1}D^\top\varepsilon \\
&= \frac{1}{n}\mathop{\mathbf{E}}_{\varepsilon}\varepsilon^\top D(D^\top D)^{-1}(D^\top D)(D^\top D)^{-1}D^\top\varepsilon && (\star) \\
&= \frac{1}{n}\mathop{\mathbf{E}}_{\varepsilon}\operatorname{Tr}\Big(\varepsilon^\top D(D^\top D)^{-1}D^\top\varepsilon\Big) && \text{(since it is a scalar)} \\
&= \frac{1}{n}\operatorname{Tr}\bigg((D^\top D)^{-1}D^\top \mathop{\mathbf{E}}_{\varepsilon}\varepsilon\varepsilon^\top D(D^\top D)^{-1}\bigg) && \text{(cyclic property)} \\
&= \frac{\sigma^2}{n}\operatorname{Tr}\Big((D^\top D)^{-1}D^\top D\Big) \\
&= \frac{\sigma^2 d}{n}
\end{aligned}
$$

In equation ($\star$), we used the fact that

$$\sum_{i=1}^{n} x^{(i)}x^{(i)\top} = D^\top D.$$

To see this, note that

$$(D^\top D)_{ij} = \sum_{k=1}^{n} x_i^{(k)}x_j^{(k)} \quad \text{and} \quad (x^{(k)}x^{(k)\top})_{ij} = x_i^{(k)}x_j^{(k)}.$$

Thus we have

$$\boxed{R(f_{LS}) = \sigma^2\left(1 + \frac{d}{n}\right).}$$

VII.4.2 Ridge Rigression

Lecture 16.
Wednesday
March 06

This time we have

$$Y = \langle w,\, X\rangle + \varepsilon$$
$$f(x; \mathcal{D}) = \langle w_{RR},\, x\rangle$$

where

$$w_{RR} = (D^\top D + \lambda I)^{-1} D^\top t.$$

We again compute the three terms:

- **Bias:** $\mathbf{E}_{\mathcal{D}|X}\, f(X; \mathcal{D}) - \langle w,\, X\rangle$ to compute $\mathbf{E}_X\Big(\mathbf{E}_{\mathcal{D}|X}\, f(X; \mathcal{D}) - \langle w,\, X\rangle\Big)^2$.
- **Variance:** $\mathrm{Var}_{\mathcal{D}|X}\, f(X; \mathcal{D})$.
- **Noise:** again simply σ^2.

Since $t = Dw + \varepsilon$, $\mathbf{E}[t] = Dw$. First we write

$$\begin{aligned} w_{RR} &= (D^\top D + \lambda I)^{-1}(D^\top D w + D^\top \varepsilon) \\ &= (D^\top D + \lambda I)^{-1}(D^\top D + \lambda I - \lambda I) w + (D^\top D + \lambda I)^{-1} D^\top \varepsilon \\ &= w - \lambda (D^\top D + \lambda I)^{-1} w + (D^\top D + \lambda I)^{-1} D^\top \varepsilon. \end{aligned}$$

Taking expectations,

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathcal{D}} w_{RR} &= w - \lambda \mathop{\mathbf{E}}_{\mathcal{D}} (D^\top D + \lambda I)^{-1} w \\ &= w - \lambda (D^\top D + \lambda I)^{-1} w. && \text{(fixed design)} \end{aligned}$$

Thus the bias is

$$\begin{aligned} \mathop{\mathbf{E}}_{\mathcal{D}} f(x; \mathcal{D}) - \langle w,\, x\rangle &= \langle x,\, \mathop{\mathbf{E}}_{\mathcal{D}} w_{RR}\rangle - \langle w,\, X\rangle \\ &= -\lambda \langle x,\, (D^\top D + \lambda I)^{-1} w\rangle. \end{aligned}$$

Lecture 16: Bias-variance decomposition with application to ridge regression

Then the expected bias^2 is

$$\begin{aligned}
\mathop{\mathbf{E}}_{X}\left(\mathop{\mathbf{E}}_{\mathcal{D}|X} f(X;\mathcal{D}) - \langle w,\, X\rangle\right)^2 &= \lambda^2 \mathop{\mathbf{E}}_{X}\left[w^\top (D^\top D + \lambda I)^{-1} x x^\top (D^\top D + \lambda I)^{-1} w\right] \\
&= \frac{\lambda^2}{n} w^\top (D^\top D + \lambda I)^{-1} \left(\sum_{i=1}^n x_i x_i^\top\right) (D^\top D + \lambda I)^{-1} w \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(fixed design)} \\
&= \frac{\lambda^2}{n} w^\top (D^\top D + \lambda I)^{-1} D^\top D (D^\top D + \lambda I)^{-1} w.
\end{aligned}$$

We can decompose $D^\top D = U\Sigma U^\top$ for some orthogonal U and diagonal Σ. Then

$$\begin{aligned}
D^\top D + \lambda I &= U(\Sigma + \lambda I)U^\top \\
(D^\top D + \lambda I)^{-1} &= U(\Sigma + \lambda I)^{-1} U^\top \\
\implies (D^\top D + \lambda I)^{-1} D^\top D (D^\top D + \lambda I)^{-1} &= U(\Sigma + \lambda I)^{-1} \Sigma (\Sigma + \lambda I)^{-1} U^\top.
\end{aligned}$$

Note that

$$(\Sigma + \lambda I)^{-1} = \operatorname{diag}\left\{\frac{1}{\sigma_i + \lambda}\right\}.$$

Thus the bias^2 term is

$$\frac{1}{n} w^\top U S U^\top w,$$

where

$$S = \operatorname{diag}\left\{\frac{\lambda^2 \sigma_i}{(\sigma_i + \lambda)^2}\right\} = \operatorname{diag}\left\{\frac{\sigma_i}{\left(1 + \frac{\sigma_i}{\lambda}\right)^2}\right\}$$

and U contains the eigenvectors of $D^\top D$.

An alternate way to write this is

$$\frac{\lambda}{n} \cdot w^\top U \left(\frac{\sqrt{\Sigma}}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{\sqrt{\Sigma}}\right)^{-2} U^\top w$$

which looks absolutely terrible.

The variance turns out to be

$$\frac{\sigma^2 d_{\text{eff}}}{n} \quad \text{where} \quad d_{\text{eff}} = \sum_{i=1}^d \frac{\sigma_i^2}{(\sigma_i + \lambda)^2}.$$

Thus we have

$$\boxed{R(f_{RR}) = \sigma^2\left(1 + \frac{d_{\text{eff}}}{n}\right) + \frac{1}{n} w^\top U S U^\top w.}$$

Chapter VIII

Maximum Likelihood Estimation

Lecture 18.
Tuesday
March 26

Definition VIII.1. Let $X_1, X_2, \dots$ be i.i.d. random variables drawn from a distribution P_θ, where θ belongs to a parameter space Θ. The *likelihood function* is defined as

$$L_n(\theta) = \prod_{i=1}^{n} \mathrm{P}_\theta(X_i),$$

which of course motivates the *log-likelihood function*

$$\ell_n(\theta) = \sum_{i=1}^{n} \log \mathrm{P}_\theta(X_i).$$

The *maximum likelihood estimator* (MLE) is defined as

$$\hat{\theta}_n = \operatorname*{argmax}_{\theta \in \Theta} \ell_n(\theta).$$

Definition VIII.2 (KL divergence)**.** The *Kullback-Leibler divergence* of the distribution P from the distribution Q is defined as

$$D_{KL}(P \parallel Q) = \mathop{\mathbf{E}}_{X \sim P}\left[\log \frac{P(X)}{Q(X)}\right].$$

For discrete distributions, this is

$$D_{KL}(P \parallel Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}.$$

Lemma VIII.3. *For all $x \in \mathbb{R}_+$,*

$$\log x \le x - 1$$

Proof. log is concave (convex down), so the tangent at $x = 1$ always lies above the curve. □

Proposition VIII.4. *For all distributions P and Q,*

$$D_{KL}(P \parallel Q) \geq 0.$$

Proof. We write $\mathbf{E}_P$ to mean $\mathbf{E}_{X \sim P}$.

$$\begin{aligned} -D_{KL}(P \parallel Q) &= \mathop{\mathbf{E}}_{P}\left[\log \frac{Q(X)}{P(X)}\right] \\ &\leq \mathop{\mathbf{E}}_{P}\left[\frac{Q(X)}{P(X)} - 1\right] \\ &= \int \frac{Q(x)}{P(x)} P(x) \mathrm{d}x - 1 \\ &= 0. \end{aligned}$$

For equality to hold, $P = Q$ almost surely. □

Exercise VIII.5. *Find the MLE of $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(\theta)$.*

Solution. The log-likelihood function is

$$\begin{aligned} \ell_n(\hat{\theta}) &= \sum_{i=1}^{n} \log \mathrm{P}_{\hat{\theta}}(X_i) \\ &= \sum_{i=1}^{n} X_i \log \hat{\theta} + (1 - X_i) \log{(1 - \hat{\theta})} \\ &= n\overline{X}_n \log \hat{\theta} + n(1 - \overline{X}_n) \log{(1 - \hat{\theta})}, \end{aligned}$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\hat{\theta}, \theta' \in [0,1]$,

$$\frac{\ell_n(\hat{\theta}) - \ell_n(\theta')}{n} = \overline{X}_n \log \frac{\hat{\theta}}{\theta'} + (1 - \overline{X}_n) \log \frac{1 - \hat{\theta}}{1 - \theta'}.$$

For $\hat{\theta} = \overline{X}_n$, this is precisely

$$\frac{\ell_n(\hat{\theta}) - \ell_n(\theta')}{n} = D_{KL}(\mathrm{Ber}\,\hat{\theta} \parallel \mathrm{Ber}\,\theta') \geq 0.$$

Thus the MLE is $\widehat{\theta}_n = \overline{X}_n$. ■

Definition VIII.6 (Entropy)**.** The *entropy* of a distribution P is defined as

$$H(P) = -\mathop{\mathbf{E}}_{X\sim P} \log P(X).$$

Exercise VIII.7. *Find MLE of $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu_0, \Sigma_0)$.*

Solution. The log-likelihood function is

$$\ell_n(\mu, \Sigma) = -\frac{1}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^{n}\|X_i - \mu\|^2_{\Sigma^{-1}} + \text{constant}.$$

Let us first fix Σ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then ignoring the constant terms, we have to maximize

$$\begin{aligned}
&\sum_{i=1}^{n}\left\|X_i - \overline{X}_n + \overline{X}_n - \mu\right\|^2_{\Sigma^{-1}} \\
&= \sum_{i=1}^{n}\left(\left\|X_i - \overline{X}_n\right\|^2_{\Sigma^{-1}} + 2\langle X_i - \overline{X}_n,\, \overline{X}_n - \mu\rangle_{\Sigma^{-1}} + \left\|\overline{X}_n - \mu\right\|^2_{\Sigma^{-1}}\right) \\
&= \sum_{i=1}^{n}\left\|X_i - \overline{X}_n\right\|^2_{\Sigma^{-1}} + 2\left\langle \cancelto{0}{\sum_{i=1}^{n} X_i - n\overline{X}_n} \,\middle|\, \Sigma^{-1} \,\middle|\, \overline{X}_n - \mu \right\rangle + n\left\|\overline{X}_n - \mu\right\|^2_{\Sigma^{-1}} \\
&= \sum_{i=1}^{n}\left\|X_i - \overline{X}_n\right\|^2_{\Sigma^{-1}} + n\left\|\overline{X}_n - \mu\right\|^2_{\Sigma^{-1}}.
\end{aligned}$$

Thus for any value of Σ, $\ell(\mu, \Sigma)$ is maximized when $\mu = \overline{X}_n$.

Now let us fix $\mu = \overline{X}_n$. Let $S = \sum_{i=1}^{n}|X_i - \mu\rangle\langle X_i - \mu|$. Then

$$\begin{aligned}
\sum\|X_i - \mu\|^2_{\Sigma^{-1}} &= \sum\langle X_i - \mu \,|\, \Sigma^{-1} \,|\, X_i - \mu\rangle \\
&= \sum \mathrm{Tr}\Big(\Sigma^{-1}|X_i - \mu\rangle\langle X_i - \mu|\Big) \\
&= \mathrm{Tr}\Big(\Sigma^{-1}S\Big).
\end{aligned}$$

Then

$$\begin{aligned}
\ell_n(\mu, \Sigma) - \ell_n(\mu, S) &\propto -\log\det\Sigma - \mathrm{Tr}\Big(\Sigma^{-1}S\Big) + \log\det S + \mathrm{Tr}\Big(S^{-1}S\Big) \\
&= \log\det\Big(\Sigma^{-1}S\Big) - \mathrm{Tr}\Big(\Sigma^{-1}S\Big) + n \\
&= \sum\log\lambda_i - \sum\lambda_i + n,
\end{aligned}$$

where λs are the eigenvalues of $\Sigma^{-1}S$.

$$
\begin{aligned}
&= \sum(\log\lambda_i - (1-\lambda_i)) \\
&\le 0
\end{aligned}
$$

with equality iff each $\lambda_i = 1$, that is, $\Sigma^{-1}S = I \iff \Sigma = S$.

Thus the MLE is

$$\widehat{\mu}_n = \overline{X}_n, \quad \widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} |X_i - \overline{X}_n\rangle\langle X_i - \overline{X}_n|. \qquad \blacksquare$$

Lecture 19. Thursday March 14

Definition VIII.8 (Consistency)**.** A sequence of estimators $\widehat{\theta}_n$ for a parameter θ is said to be *consistent* if

$$\widehat{\theta}_n \xrightarrow{\mathrm{P}} \theta.$$

That is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathrm{P}(|\widehat{\theta}_n - \theta| < \epsilon) = 1.$$

Theorem VIII.9 (Central limit theorem)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance σ^2. Then*

$$\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Theorem VIII.10. *Let $\{f_\theta(x) \mid \theta \in \Theta\}$ be a nice family of distributions. Let X_is be distributed according to the $f_{\theta_0}(x)$. Then the MLE $\widehat{\theta}_n$ is consistent.*

Proof. By the law of large numbers,

$$\frac{1}{n}\ell_n(\theta) \xrightarrow{\mathrm{P}} \mathbf{E}[\log f_\theta(X_i)] \quad \text{if it exists.}$$

This expectation is under the true parameter θ_0.

Thus

$$
\begin{aligned}
\frac{1}{n}\ell_n(\theta) - \frac{1}{n}\ell_n(\theta_0) &\xrightarrow{\mathrm{P}} \mathop{\mathbf{E}}_{\theta_0}\left[\log\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right] \\
&= D_{KL}(f_{\theta_0} \parallel f_\theta) \\
&\ge 0,
\end{aligned}
$$

where the equality holds iff $f_{\theta_0}(X) = f_\theta(X)$ almost everywhere. Thus

$$\operatorname*{argmax}_\theta \ell_n(\theta) \to \theta_0. \qquad \square$$

Lecture 19: Consistency, Normality, Efficiency and Bias

VIII.1 How Good is This Convergence?

In this section we define

Definition VIII.11 (Log-likelihood)**.** The *normalized log-likelihood* function is

$$L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i).$$

for convenience.

Definition VIII.12 (Score function)**.** Given a probability mass/density family $f_\theta(x)$, where $\theta \in \mathbb{R}^d$ is the parameter, the *score function* $Z_x \colon \Theta \subseteq \mathbb{R}^d \to \mathbb{R}^d$ is defined by

$$Z_x(\theta) = \frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{1}{f_\theta(x)} \frac{\partial}{\partial\theta} f_\theta(x).$$

We will also abuse the heck out of notation and write

$$Z_x'(\theta) \text{ to mean } \frac{\partial}{\partial\theta} Z_x(\theta),$$

which is reasonable, but

$$f_\theta'(x) \text{ to mean } \frac{\partial}{\partial\theta} f_\theta(x),$$

which is not. To make it seem less unreasonable, we may write

$$f_\theta(x) \text{ as } f(x \mid \theta).$$

Lemma VIII.13. *For $\theta \in \Theta$,*

$$\boldsymbol{E}[Z_X(\theta)] = 0, \quad \operatorname{Var}[Z_X(\theta)] = -\boldsymbol{E}[Z_X'(\theta)].$$

The expectations and variances are over $X \sim f(\theta)$. First of all, "For $\theta \in \Theta$" specifies that θ is fixed. Secondly, we will never in this section prescribe a distribution over Θ. That would be a weird thing to do.

Proof.

$$\begin{aligned}
\mathbf{E}[Z_X(\theta)] &= \int \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{\cancel{f_\theta(x)}} \cancel{f_\theta(x)} \mathrm{d}x \\
&= \frac{\partial}{\partial\theta} \int f_\theta(x) \mathrm{d}x \\
&= 0.
\end{aligned}$$

Alternatively,

$$\begin{aligned}
\mathbf{E}[Z_X(\theta)] &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_{X_i}(\theta) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(X_i) \\
&= \frac{\partial}{\partial\theta} \lim_{n\to\infty} L_n(\theta) \\
&= 0.
\end{aligned}$$

(Something of that sort.)

Next note that

$$\begin{aligned}
Z_X'(\theta) &= \frac{\partial}{\partial\theta} Z_X(\theta) \\
&= \frac{\partial}{\partial\theta} \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)} \\
&= \frac{\frac{\partial^2}{\partial\theta^2} f_\theta(x)}{f_\theta(x)} - \frac{\left(\frac{\partial}{\partial\theta} f_\theta(x)\right)^2}{f_\theta(x)^2} \\
&= \frac{\frac{\partial^2}{\partial\theta^2} f_\theta(x)}{f_\theta(x)} - Z_X(\theta)^2.
\end{aligned}$$

The expectation of the first term is again 0 by the same reasoning. Thus

$$\operatorname{Var}[Z_X(\theta)] = \mathbf{E}[Z_X(\theta)^2] = -\mathbf{E}[Z_X'(\theta)]. \qquad \square$$

Theorem VIII.14. *Let $\{f_\theta(x) \mid \theta \in \Theta\}$ be a family of distributions and assume all niceties. Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. random variables distributed according to f_{θ_0}. Let $\widehat{\theta}_n$ be the MLE of θ based on $X_1, \dots, X_n$. Then $\widehat{\theta}_n$ is consistent and asymptotically normal, with*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{\mathcal{I}(\theta_0)}\right),$$

where

$$\mathcal{I}(\theta) = \operatorname*{Var}_{X\sim f(\theta)}[Z_X(\theta)] = -\operatorname*{\boldsymbol{E}}_{X\sim f(\theta)}[Z_X'(\theta)]$$

is the Fisher information matrix*.*

$\frac{1}{\mathcal{I}(\theta_0)}$ obviously refers to the inverse of the matrix.

"Proof". Assume that Z_x is thrice differentiable for each x. Then

$$Z_x(\theta) = Z_x(\theta_0) + Z'_x(\theta_0)(\theta - \theta_0) + \frac{1}{2}Z''_x(\theta_0)(\theta - \theta_0)^2 + O(\|\theta - \theta_0\|^3).$$

Summing over all x's,

$$\sum_{i=1}^{n} Z_{X_i}(\theta) \approx \sum_{i=1}^{n} Z_{X_i}(\theta_0) + \sum_{i=1}^{n} Z'_{X_i}(\theta_0)(\theta - \theta_0)$$

This is relevant because

$$L'_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} Z_{X_i}(\theta).$$

Now note that

$$\begin{aligned} L'_n(\theta) &= L'_n(\theta_0) + L''_n(\theta_0)(\theta - \theta_0) + O(\|\theta - \theta_0\|^2) \\ \implies 0 &= L'_n(\theta_0) + L''_n(\theta_0)(\widehat{\theta}_n - \theta_0) + O(\left\|\widehat{\theta}_n - \theta_0\right\|^2), \end{aligned}$$

since $\widehat{\theta}_n$ maximizes L_n. Since $\widehat{\theta}_n$ is consistent, the second order term can be ignored. Thus

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \approx -\sqrt{n}L'_n(\theta_0)L''_n(\theta_0)^{-1}.$$

Now

$$\sqrt{n}L'_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{X_i}(\theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)),$$

since the Z_{X_i} are i.i.d. with mean 0 and variance $\mathcal{I}(\theta_0)$.

Moreover

$$\begin{aligned} L''_n(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n} Z'_X(\theta_0) \\ &\xrightarrow{\mathrm{P}} -\mathbf{E}[Z'_X(\theta_0)] = \mathcal{I}(\theta_0). \end{aligned}$$

Wow! Look at the interplay between the variance of the score, and the expectation of its derivative.

This gives (appealing to intuition)

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \frac{1}{\mathcal{I}(\theta_0)}N(0, \mathcal{I}(\theta_0)) = N(0, \mathcal{I}(\theta_0)^{-1}).$$ □

Definition VIII.15 (Mean squared error)**.** The *mean squared error* of an estimator $T_n \colon \mathcal{X}^n \to \Theta$ for the parameter θ_0 is

$$\mathrm{MSE}(T_n) = \mathbf{E}[(T_n - \theta_0)^2].$$

Among two unbiased estimators, the one with the lower variance is deemed to be better.

VIII.2 Efficiency & Bias

We now attempt to justify why MLE is a good choice, apart from proof by obviousness. In fact, it is not obvious, because there are estimators which have a lower MSE than the MLE. Which is better in such a case?

Theorem VIII.16 (Cramér-Rao bound)**.** *Let $T_n \colon \mathcal{X}^n \to \Theta$ be an unbiased estimator for θ using i.i.d. samples $X_1, \dots, X_n$. Then*

$$\operatorname{Var}(T_n) \geq \frac{1}{n\mathcal{I}(\theta_0)}.$$

Proof. Let $Z(\theta) = \ell_n'(\theta) = \sum_{i=1}^n Z_{X_i}(\theta)$. Obviously $\operatorname{Var}[Z(\theta)] = \sum \operatorname{Var}(Z_{X_i}(\theta)) = n\mathcal{I}(\theta)$.

Also note that $\mathbf{E}[Z(\theta)] = 0$.

Now since T_n is unbiased,

$$\theta = \mathbf{E}[T_n \mid \theta]$$

for all $\theta \in \Theta$. Differentiating both sides,

$$\begin{aligned} 1 &= \frac{\partial}{\partial\theta}\mathbf{E}[T_n \mid \theta] \\ &= \frac{\partial}{\partial\theta}\int T_n(\boldsymbol{x}) f_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int T_n(\boldsymbol{x})\frac{\partial}{\partial\theta} f_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \int T_n(\boldsymbol{x}) Z(\theta)(\boldsymbol{x}) f_\theta(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= \mathbf{E}[T_n Z(\theta)] \\ &= \operatorname{Cov}(T_n, Z(\theta)). \end{aligned}$$

What?! Two random variables that are perfectly correlated!

Now

$$\begin{aligned} \operatorname{Cov}(T_n, Z(\theta)) &\leq \sqrt{\operatorname{Var}(T_n)\operatorname{Var}(Z(\theta))} \\ \implies \operatorname{Var}(T_n) &\geq \frac{\operatorname{Cov}(T_n, Z(\theta))^2}{\operatorname{Var}(Z(\theta))} \\ &= \frac{1}{n\mathcal{I}(\theta)}. \end{aligned}$$

□

Lecture 19: Consistency, Normality, Efficiency and Bias

But from theorem VIII.14, we already know that the MLE achieves this bound. Thus the MLE is *asymptotically efficient*.

Chapter IX

EM Algorithm

IX.1 Latent Variable Models

Lecture 20.
Monday
March 18

Let $\{P_\theta \mid \theta \in \Theta\}$ be a family of distributions, where $P_\theta \colon \mathcal{X} \times \mathcal{Z} \to [0, \infty)$. Let $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$ be random variables with joint distribution P_θ. That is,

$$\mathrm{P}(X = x, Z = z \mid \theta) = P_\theta(x, z).$$

So

$$\mathrm{P}(X = x \mid \theta) = \sum_z P_\theta(x, z)$$

X is the *observed* variable and Z is the *latent* variable.

Let

$$\mathcal{D} = \left\{ x^{(1)}, \ldots, x^{(N)} \right\} \subseteq \mathbb{R}^n.$$

For a fixed $x \in \mathcal{X}$, define the posterior

$$\mathrm{Pos}(\theta) = z \mapsto \mathrm{P}(Z = z \mid X = x, \theta).$$

Definition IX.1. Given $\boldsymbol{x} = (x^{(i)})_{i=1}^N$, define the *complete data likelihood* as

$$f(\boldsymbol{x}, \boldsymbol{Z} \mid \theta) = \prod_{i=1}^{N} P_\theta(x^{(i)}, Z^{(i)}),$$

where $\boldsymbol{Z} = (Z^{(i)})_{i=1}^N$ is a random vector over $\mathcal{Z}^N$, and $\theta \in \Theta$.

Use this to define

$$Q(\theta, \theta^{(0)}) = \mathbf{E}[\log f(\boldsymbol{x}, \boldsymbol{Z} \mid \theta)]$$

This is the *expected complete data log-likelihood.*

Proposition IX.2. *If* $\mathcal{Z} = [k]$, *then*

$$Q(\theta, \theta^{(0)}) = \sum_{j=1}^{N} Q^{(j)}(\theta, \theta^{(0)})$$

where

$$Q^{(j)}(\theta, \theta^{(0)}) = \sum_{i=1}^{k} P_i^{(j)} \log P_\theta(x^{(j)}, i)$$

and

$$P_i^{(j)} = \mathrm{P}(Z^{(j)} = i \mid X^{(j)} = x^{(j)}, \theta^{(0)}).$$

Proof. Expand the expectation.

$$\begin{aligned}
Q(\theta, \theta^{(0)}) &= \mathbf{E}\left[\sum_{j=1}^{N} \log P_\theta(x^{(i)}, Z_i) \,\middle|\, X^{(j)} = x^{(j)}, \theta^{(0)}\right] \\
&= \sum_{j=1}^{N} \mathbf{E}[\log P_\theta(x^{(i)}, Z_i) \mid X^{(j)} = x^{(j)}, \theta^{(0)}] \\
&= \sum_{j=1}^{N} \sum_{i=1}^{k} \mathrm{P}(Z^{(j)} = i \mid X^{(j)} = x^{(j)}, \theta^{(0)}) \log P_\theta(x^{(j)}, i) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{k} P_i^{(j)} \log P_\theta(x^{(j)}, i) \\
&= \sum_{j=1}^{N} Q^{(j)}(\theta, \theta^{(0)}).
\end{aligned}$$

□

But we don't even see the $Z^{(i)}$'s! What we wish to maximize is the likelihood of the observed data

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{N} \log \mathrm{P}(X = x^{(i)} \mid \theta) \\
&= \sum_{i=1}^{N} \log \sum_{z \in \mathcal{Z}} P_\theta(x^{(i)}, z) \\
&= \sum_{i=1}^{N} \log \mathop{\mathbf{E}}_{Z \sim \mathrm{Pos}(\theta)} [P_\theta(x^{(i)}, Z)].
\end{aligned}$$

The KL divergence of P_θ from $P_{\theta'}$ is

$$\begin{aligned} D_{KL}(P_\theta \parallel P'_\theta) &= \mathop{\mathbf{E}}_{X,Z\sim P_\theta}[\log P_\theta(X,Z) - \log P_{\theta'}(X,Z)] \\ &= \mathop{\mathbf{E}}_{X\sim P_\theta^X} \mathop{\mathbf{E}}_{Z\sim \mathrm{Pos}(\theta)}[\log P_\theta(X,Z) - \log P_{\theta'}(X,Z)]. \end{aligned}$$

IX.2 Restricted Boltzmann Machines

Lecture 21. Tuesday March 26

$$\mathrm{P}(S = s) = \frac{e^{-E(s)/T}}{Z}$$

where T is the temperature. We will fix $T = 1$.

$$\mathrm{P}(S = s) = \frac{e^{-E(s)}}{Z}$$

Suppose $w_{ii} = 0$ and (w_{ij}) is symmetric.

$$\begin{aligned} E(s) &= -\frac{1}{2}\sum_{i,j} w_{ij}s_is_j - \sum_i b_is_i \\ &= \sum_i s_i\Big(\sum_{j\geq i} w_{ij}s_j + b_i\Big) \\ &= \sum_i s_i\Big(\sum_{j> i} w_{ij}s_j + b_i\Big) \end{aligned}$$

since $w_{ii} = 0$. Suppose further that $w_{12} = 0$. Then

$$E(s) = s_1\Big(\sum_{j>2} w_{1j}s_j + b_1\Big) + s_2\Big(\sum_{j>2} w_{2j}s_j + b_2\Big) + K$$

where K depends only on $s_3, \ldots, s_n$. Thus conditioned on $S_{3:n}$, S_1 and S_2 are conditionally independent.

IX.2.1 A real life example

You feel sick. You go to the doctor. The doctor asks you a series of questions, perhaps about the weather, your kids, your job, your symptoms. The doctor then diagnoses you. The doctor is a restricted Boltzmann machine?

The doctor has a knowledge base

$$\mathrm{P}(S = s \mid D_1 = d_1, \ldots, D_m = d_m).$$

They figure out the inverse of this,

$$\mathrm{P}(D = d \mid S = s)$$

using some Bayesion wizardry.

Now that we have touched some grass, let's go back to the math.

Split $S = (V, H)$ where V are the "visible" symptoms and H are the "hidden" symptoms.

$$V = \begin{pmatrix} S_1 \\ \vdots \\ S_m \end{pmatrix}, \quad H = \begin{pmatrix} S_{m+1} \\ \vdots \\ S_d \end{pmatrix}$$

Let

$$\mathcal{D} = \left\{ v^{(1)}, \ldots, v^{(N)} \right\}$$

We apply the latent variable model.

$$\log \mathrm{P}(V = v) = \log \sum_h \mathrm{P}(V = v, H = h)$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log \mathrm{P}(V = v^{(i)})$$

and we employ the algorithm

$$\theta \leftarrow \theta + \eta \frac{\partial \mathcal{L}}{\partial \theta}$$

where η is the learning rate. This is called *gradient ascent*.

Consider the baby case $N = 1$.

$$
\begin{aligned}
\mathcal{L} &= \log \sum_h \mathrm{P}(V = v, H = h) \\
&= \log \sum_h e^{-E(s)} - \log Z \\
\implies \frac{\partial \mathcal{L}}{\partial w_{ij}} &= \frac{1}{\sum_h e^{-E(s)}} \sum_h e^{-E(s)} \frac{\partial E(s)}{\partial w_{ij}} - \frac{1}{Z} \frac{\partial Z}{\partial w_{ij}} \\
&= \frac{1}{\sum_h e^{-E(s)}} \sum_h e^{-E(s)} \frac{\partial E(s)}{\partial w_{ij}} + \frac{1}{Z} \sum_s e^{-E(s)} \frac{\partial E(s)}{\partial w_{ij}} \\
&= \sum_h \frac{e^{-E(s)}}{\sum_h e^{-E(s)}} s_i s_j - \sum_s \frac{e^{-E(s)}}{Z} s_i s_j \\
&= \sum_h \mathrm{P}(H = h \mid V = v) s_i s_j - \sum_s \mathrm{P}(S = s) s_i s_j \\
&= \mathop{\mathbf{E}}_{H|V}[s_i s_j] - \mathop{\mathbf{E}}_{S}[s_i s_j]
\end{aligned}
$$

Lecture 24. Thursday April 04

Let $(Z_t)_{t \in \mathbb{N}}$ be a Markov chain with state space $S = \{s_1, \ldots, s_n\}$ and transition matrix $A = (a_{ij})_{i,j=1}^n$. That is,

$$
\mathrm{P}(Z_{t+1} = s_j \mid Z_t = s_i) = a_{ij}.
$$

If $Z_0 \sim \mu^\top$, then the distribution of Z_n is $\mu^\top A^n$.

The Ising model is incredibly hard to compute. Instead of computing

$$
\mathbf{E}[S_i] = \sum_s s_i \, \mathrm{P}(S = s),
$$

we can sample

$$
\mathbf{E}[S_i] \approx \frac{1}{m} \sum_{i=1}^m s_i^*,
$$

where s_i^*s are sampled according to the distribution $\mathrm{P}(S = \cdot)$.

This is done via the Metropolis algorithm.

Chapter X

Graphical Models

Let $\{X_i\}_{i=1}^d$ be a set of random variables. To each X_i assign a vertex i, and let the vertex set be $[d]$. Edges will model dependencies between the random variables.

We first review some definitions from graph theory.

X.1 Definitions

Definition X.1. A *graph* $G = (V, E)$ consists of a set of vertices V and a set of edges $E \subseteq V \times V$.

A graph is *undirected* if $(u, v) \in E$ implies $(v, u) \in E$. Otherwise, it is *directed*.

A vertex v is *adjacent* to u if $(u, v) \in E$. u is said to be the *parent* of v.

The *neighbourhood* of v is the set of vertices adjacent to v.

$$N(v) = \{u \in V \mid (u, v) \in E\}.$$

Definition X.2 (Paths)**.** A *path* in a graph $G = (V, E)$ is a sequence of vertices $(v_1, \ldots, v_k)$ such that $(v_i, v_{i+1}) \in E$ for all i.

A path is *simple* if all vertices are distinct. A path is *closed* if $v_1 = v_k$.

A *cycle* is a closed path with no repeated vertices (except for the first and last).

Definition X.3 (Separation)**.** Let $A, B, C \subseteq V$ be disjoint. A and B are *separated* by C if every path from A to B contains a vertex in C.

We now study two instances of graphical models:

- Bayesian networks
- Markov networks

X.1.1 Bayesian Networks

Definition X.4. Let $G = ([n], E)$ be a directed acyclic graph. Then $(G, X_1, \dots, X_n)$ is a *Bayesian network* if

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_i \mid \text{parent}(i))$$

where parent(i) is the set of parents of i.

For convenience, we will relabel the vertices in a topological sort. Then for all i,

$$\text{parent}(i) \subseteq [i-1]$$

X.2 A real life example

Let N, T, L, X be random variables representing the following:

- N represents whether a particular patient has pneumonia.
- T represents whether they have tuberculosis.
- L represents whether they have observable lung abnormalities.
- X represents whether they have a positive X-ray.

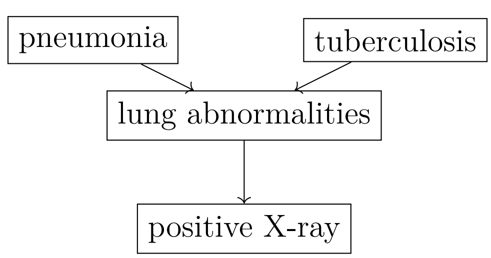
Then

$$\begin{aligned}&\mathrm{P}(N = n, T = t, L = l, X = x)\\ &\qquad = \mathrm{P}(N = n)\,\mathrm{P}(T = t)\,\mathrm{P}(L = l \mid N = n, T = t)\,\mathrm{P}(X = x \mid L = l).\end{aligned}$$

We will shorten such equations to

$$\mathrm{P}(N, T, L, X) = \mathrm{P}(N)\,\mathrm{P}(T)\,\mathrm{P}(L \mid N, T)\,\mathrm{P}(X \mid L).$$

This can be represented by the following Bayesian network:



Lecture 24

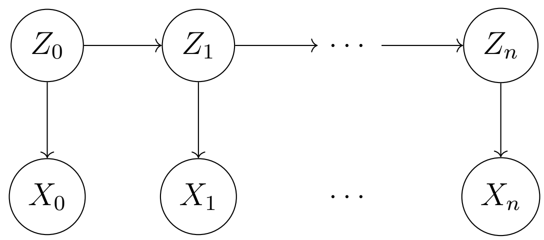
The *inference problem* is to compute

$$\mathrm{P}(N = 1 \mid X = x).$$

Suppose the X-ray machines are awesome, so that $L = X$ with probability 1.

X.2.1 HMMs as Bayesian Networks

Consider the following Bayesian network:



The total probability is

$$\mathrm{P}(X_0, Z_0, \ldots, X_n, Z_n) = \mathrm{P}(Z_0) \prod_{i=1}^{n} \mathrm{P}(Z_i \mid Z_{i-1}) \prod_{i=0}^{n} \mathrm{P}(X_i \mid Z_i).$$

This is the hidden Markov model!

X.3 Markov Networks

Definition X.5 (Global Markov property)**.** Let $G = ([n], E)$ be undirected. Then $(G, X_1, \ldots, X_n)$ satisfies the *global Markov property* if for all $A, B, C \subseteq [n]$ such that A and B are separated by C,

$$X_A \perp\!\!\!\perp X_B \mid X_C,$$

where $X_S = \{X_i\}_{i \in S}$.

Theorem X.6 (Hammersly-Clifford theorem)**.** *If $(G, X_1, \ldots, X_n)$ satisfies the global Markov property, and $P(X_1, \ldots, X_n) > 0$, then the joint distribution of $X_1, \ldots, X_n$ factorizes over G. That is,*

$$\mathrm{P}(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{C \in cliques(G)} \psi_C(X_C),$$

where Z is a normalizing constant and ψ_C is a potential function.

Chapter XI

Principal Component Analysis

Lecture 25.
Monday
April 08

Let the data

$$\mathcal{D} = \left\{x^{(1)}, \ldots, x^{(n)}\right\} \subseteq \mathbb{R}^d$$

be drawn i.i.d. from a distribution P.

By Cauchy-Shwartz,

$$\langle u, v\rangle \leq \|u\|\|v\|,$$

with equality achieved when $v = \lambda u$.

THEREFORE, the maximum value of $u^\top C u$ is achieved... when $Cu = \lambda u$.

This reminds me of

$$E = mc^2$$

$$E = \frac{hc}{\lambda}$$

$$\lambda = \frac{h}{mv}$$