Feb 7 Notes (Session 1)

Gaussian Distribution: (some plots for different values of \mu and \sigma)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Joint Distribution:

Two coins might be both fair, but probability of them both being 1 at the same time can be 1/8 (as opposed to ¼). That is "joint" distribution.

Bivariate Gaussian Distribution: (define what \mu, \sigma and \rho are)

$$\boxed{\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \qquad \textbf{(Eq.2)}}$$

encodes the correlation between X and Y

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right)$$

Special case where \rho = 0 (they become independent)

Demonstrate the impacts of each variable on the distribution plot using:

https://demonstrations.wolfram.com/TheBivariateNormalDistribution/

If we define

$$\boldsymbol{\mu} = \begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}$$

Then,

$$\boxed{f_{\mathbf{X}}(x_1,\ldots,x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}}$$

Which is **Multivariate Gaussian Distribution**

with k-dimensional mean vector

$$\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}] = (\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]),$$

and $k \times k$ covariance matrix

$$\Sigma_{i,j} := \mathrm{E}[(X_i-\mu_i)(X_j-\mu_j)] = \mathrm{Cov}[X_i, X_j]$$

Theorem: \Sigma is Positive Semi-Definite.

Proof: Covariance matrix **C** is calculated by the formula,

$$\mathbf{C} \triangleq E\{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\}$$

For an arbitrary real vector **u**, we can write,

$$\mathbf{u}^T\mathbf{C}\mathbf{u} = \mathbf{u}^T E\{(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\}\mathbf{u} = E\{\mathbf{u}^T(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})^T\mathbf{u}\} = E\{z^2\} >= 0.$$

Maximum Likelihood Estimation: given some data, how do we estimate \mu and \Sigma? Generally, given some observations D, how do we estimate parameter \Theta?

$X_1, X_2, X_3, \ldots X_n$ have joint density denoted

$$f_\theta(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta)$$

Given observed values $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the likelihood of θ is the function

$$lik(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$$

considered as a function of θ.

If the distribution is discrete, f will be the frequency distribution function.

In words: $\boxed{\text{lik}(\theta)\text{=probability of observing the given data as a function of } \theta.}$

Definition:

The maximum likelihood estimate (mle) of θ is that value of θ that maximises $lik(\theta)$: it is the value that makes the observed data the "most probable".

If the X_i are iid, then the likelihood simplifies to

$$lik(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Rather than maximising this product which can be quite tedious, we often use the fact that the logarithm is an increasing function so it will be equivalent to maximise the log likelihood:

$$l(\theta) = \sum_{i=1}^{n} \log(f(x_i|\theta))$$

Normal example

If $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ random variables their density is written:

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_i^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2\right)$$

Regarded as a function of the two parameters, μ and σ this is the likelihood:

$$\ell(\mu, \sigma) = -nlog\sigma - \frac{n}{2}log2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial \ell}{\partial \sigma} = -\tfrac{n}{\sigma} + \sigma^{-3}\textstyle\sum_{i=1}^{n}(x_i - \mu)^2$$

so setting these to zero gives $\bar{X}$ as the mle for μ, and $\hat{\sigma}^2$ as the usual.

MLE is NOT always unbiased, e.g. \sigma^2 needs to be divided by "n-1".

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\mathbb{E}[(x_j - \hat{\mu})^2] \\ &= \mathbb{E}[x_j^2] - 2\mathbb{E}[x_j\hat{\mu}] + \mathbb{E}[\hat{\mu}^2] \\ &= \sigma^2 + \mu^2 - 2\left(\tfrac{n-1}{n}\mu^2 + \tfrac{1}{n}(\sigma^2 + \mu^2)\right) + \left(\tfrac{n^2-n}{n^2}\mu^2 + \tfrac{n}{n^2}(\sigma^2 + \mu^2)\right) \\ &= \tfrac{n-1}{n}\sigma^2 \end{aligned}$$

$$\mathbb{E}[x_j x_k] = \begin{cases} \mathbb{E}[x_j]\mathbb{E}[x_k] = \mu^2 & \text{if } j \neq k \\ \mathbb{E}[x_j^2] = \sigma^2 + \mu^2 & \text{if } j = k \end{cases}$$

Remember the expected value of ***x_i²*** mentioned at the start? By expanding ^***μ***, we have

$$\begin{aligned} \mathbb{E}[x_j\hat{\mu}] &= \tfrac{n-1}{n}\mu^2 + \tfrac{1}{n}(\sigma^2 + \mu^2) \\ \mathbb{E}[\hat{\mu}^2] &= \tfrac{n^2-n}{n^2}\mu^2 + \tfrac{n}{n^2}(\sigma^2 + \mu^2) \end{aligned}$$

Feb 7 Notes (Session 2)

Regression Fit/Overfit: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-slides-1.pdf

Linear Models:
Polynomial curve fitting:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Basis Functions:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Typically first basis function is just bias term.
Identity basis functions
Polynomial basis functions

$$\phi_j(x) = x^j.$$

Gaussian basis functions

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

Sigmoid basis functions

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

where

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

Generalized Linear Models: y = f(w^T \phi(x))

MLE & MSE relationship

Maximum Likelihood and Least Squares (1)

Assume observations from a deterministic function with added Gaussian noise:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \quad \text{where} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

which is the same as saying,

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$, we obtain the likelihood function

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

Maximum Likelihood and Least Squares (2)

Taking the logarithm, we get

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

where

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

is the sum-of-squares error.

Maximum Likelihood and Least Squares (3)

Computing the gradient and setting it to zero yields

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \mathbf{0}.$$

Solving for $\mathbf{w}$, we get

$$\mathbf{w}_{\mathrm{ML}} = \underbrace{\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}}_{\text{The Moore-Penrose pseudo-inverse, } \boldsymbol{\Phi}^{\dagger}.}\mathbf{t}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Geometry of Least Squares

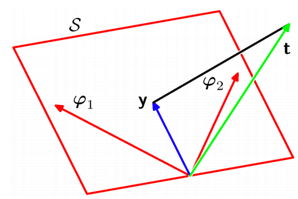
Consider

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M]\,\mathbf{w}_{\mathrm{ML}}.$$

$$\mathbf{y} \in \underbrace{\mathcal{S}}_{M\text{-dimensional}} \subseteq \underbrace{\mathcal{T}}_{N\text{-dimensional}} \qquad \mathbf{t} \in \mathcal{T}$$

$\mathcal{S}$ is spanned by $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$.

$\mathbf{w}_{\mathrm{ML}}$ minimizes the distance between $\mathbf{t}$ and its orthogonal projection on $\mathcal{S}$, i.e. $\mathbf{y}$.



Regularization

Regularized Least Squares (1)

Consider the error function:

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

With the sum-of-squares error function and a quadratic regularizer, we get

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

which is minimized by

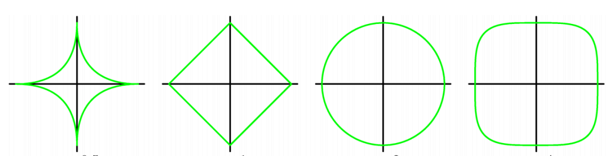
$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}.$$

λ is called the regularization coefficient.

Regularized Least Squares (2)

With a more general regularizer, we have

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

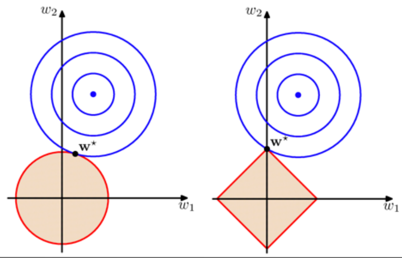


$q = 0.5$ $q = 1$ $q = 2$ $q = 4$

Lasso Quadratic

Regularized Least Squares (3)

Lasso tends to generate sparser solutions than a quadratic regularizer.



Multiple Output (Multi-task Learning)

Multiple Outputs (1)

Analogously to the single output case we have:

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) &= \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{W}, \mathbf{x}), \beta^{-1}\mathbf{I}) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I}). \end{aligned}$$

Given observed inputs, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and targets, $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_N]^{\mathrm{T}}$, we obtain the log likelihood function

$$\begin{aligned} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2} \ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2} \sum_{n=1}^{N} \left\|\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\|^2. \end{aligned}$$

Multiple Outputs (2)

Maximizing with respect to $\mathbf{W}$, we obtain

$$\mathbf{W}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{T}.$$

If we consider a single target variable, t_k, we see that

$$\mathbf{w}_k = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}_k = \boldsymbol{\Phi}^{\dagger}\mathbf{t}_k$$

where $\mathbf{t}_k = [t_{1k}, \ldots, t_{Nk}]^{\mathrm{T}}$, which is identical with the single output case.