[redacted]

QUESTION 1

Given a soft margin linear SVM problem

$$\max_{\mathbf{0}\leq\boldsymbol{\alpha}\leq C} f(\boldsymbol{\alpha}) \quad = \quad \boldsymbol{\alpha}^T\mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{G}\boldsymbol{\alpha}$$

where all the vectors and matrices have their usual meanings as discussed in class. The given problem handles optimization process using a variant of coordinate gradient descent based approach. The updation rule for this approach is as follows:

$$\begin{aligned}\alpha_n &= \alpha_n + \delta_* \\ \delta* &= \arg\max_{\delta} f(\boldsymbol{\alpha} + \delta\mathbf{e}_n)\end{aligned}$$

where $\mathbf{e}_n$ denotes a vector of all zeros except a 1 entry at n.

$$\begin{aligned}f(\boldsymbol{\alpha} + \delta\mathbf{e}_n) &= (\boldsymbol{\alpha} + \delta\mathbf{e}_n)^T\mathbf{1} - \frac{1}{2}(\boldsymbol{\alpha} + \delta\mathbf{e}_n)^T\mathbf{G}(\boldsymbol{\alpha} + \delta\mathbf{e}_n) \\ &= f(\boldsymbol{\alpha}) + \delta(1 - \mathbf{e}_n^T\mathbf{G}\boldsymbol{\alpha}) - \frac{1}{2}\delta^2\mathbf{e}_n^T\mathbf{G}\mathbf{e}_n\end{aligned}$$

Maximizing w.r.t. to δ we have:

$$\begin{aligned}\frac{\partial}{\partial\delta}f(\boldsymbol{\alpha} + \delta\mathbf{e}_n) &= 1 - \mathbf{e}_n^T\mathbf{G}\boldsymbol{\alpha} - \delta\mathbf{e}_n^T\mathbf{G}\mathbf{e}_n = 0 \\ \delta^* &= \frac{(1 - \mathbf{e}_n^T\mathbf{G}\boldsymbol{\alpha})}{\mathbf{e}_n^T\mathbf{G}\mathbf{e}_n}\end{aligned}$$

Simplifying further we have

$$\delta^* = \frac{1 - y_n\mathbf{w}^T\mathbf{x}_n}{\|\mathbf{x}_n\|^2}$$

$\because \ \mathbf{e}_n^T\mathbf{G}\boldsymbol{\alpha} = \sum_{m=1}^{N} y_n y_m \mathbf{x}_n^T\mathbf{x}_m\alpha_m = y_n x_n^T\big(\sum_{m=1}^{N} y_m\mathbf{x}_m\alpha_m\big) = y_n\mathbf{w}^T\mathbf{x}_n$ and $\mathbf{e}_n^T\mathbf{G}\mathbf{e}_n = y_n^2\|\mathbf{x}_n\|^2 = \|\mathbf{x}_n\|^2$

The above method doesn't put a constraint on updated α_n. To handle this we use projected gradient descent i.e if $\alpha_n + \delta^* \in [0, C]$ we update in a regular fashion else if $\alpha_n + \delta^* < 0$ we update α_n to 0 else if $\alpha_n + \delta^* > C$ we update α_n to C.

Algorithm 1 Co-ordinate Ascent for Soft Margin SVM

1. *Initialize* $\boldsymbol{\alpha}$ *as* $\hat{\boldsymbol{\alpha}}$ *s.t* $0 < \alpha_n < C \ \forall \ n = 1, \ldots, N$
2. *For* i *in* $[1, \ldots, N]$

$$\alpha_i = \begin{cases}\alpha_i + \delta^* & \alpha_i + \delta^* \in [0, C] \\ 0 & \alpha_i + \delta^* < 0 \\ C & \alpha_i + \delta^* > C\end{cases}$$

3. *If not converged go to step* 2

Ridge Regression: Revisited

Question 2

- Recall the ridge regression problem

$$\boldsymbol{w} = \arg\min_{\boldsymbol{w}} \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \lambda \boldsymbol{w}^\top \boldsymbol{w}$$

- The solution to this problem was

$$\boldsymbol{w} = (\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top + \lambda \mathbf{I}_D)(\sum_{n=1}^{N} y_n \boldsymbol{x}_n) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_D)^{-1} \mathbf{X}^\top \boldsymbol{y}$$

- Inputs don't appear as inner-products here . They actually do! :-)
- Matrix inversion lemma: $(\mathbf{F}\mathbf{H}^{-1}\mathbf{G} - \mathbf{E})^{-1}\mathbf{F}\mathbf{H}^{-1} = \mathbf{E}^{-1}\mathbf{F}(\mathbf{G}\mathbf{E}^{-1}\mathbf{F} - \mathbf{H})^{-1}$
- The lemma allows us to rewrite $\boldsymbol{w}$ as

$$\boldsymbol{w} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)^{-1} \boldsymbol{y} = \mathbf{X}^\top \boldsymbol{\alpha} = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}_n$$

  where $\boldsymbol{\alpha} = (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N)^{-1} \boldsymbol{y} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1} \boldsymbol{y}$ is an $N \times 1$ vector of dual variables, and $K_{nm} = \boldsymbol{x}_n^\top \boldsymbol{x}_m$

- Note: $\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \boldsymbol{x}_n$ is known as "dual" form of ridge regression solution. However, so far it is still a linear model. But now it is easily kernelizable.

Question 3

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\Big\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big\} \\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\Big\{-\frac{1}{2}\text{trace}\Big[\boldsymbol{\Sigma}^{-1}\mathbf{S}\Big]\Big\} \qquad \text{where } \mathbf{S}=(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\Big\{-\frac{1}{2}\Big(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\Big)\Big\}$$

  where $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are the "natural parameters" (recall exp. family).

- Note that there is a term quadratic in $\boldsymbol{x}$ (involves $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$) and linear in $\boldsymbol{x}$ (involves $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$)
- Information form can help recognize $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a Gaussian when doing algebraic manipulations

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$ and ignoring the constants

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) \\
&= \frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}\text{trace}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^\top] \\
&= \frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\text{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}_\mu] \qquad \left[\text{where} \quad \mathbf{S}_\mu = \sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^\top\right]
\end{aligned}$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right] = -\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}} \left[\frac{N}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \text{trace}[\boldsymbol{\Lambda} \mathbf{S}_\mu] \right] = \frac{N}{2} \boldsymbol{\Lambda}^{-\top} - \frac{1}{2} \mathbf{S}_\mu^\top = \frac{N}{2} \boldsymbol{\Lambda}^{-1} - \frac{1}{2} \mathbf{S}_\mu = \frac{N}{2} \boldsymbol{\Sigma} - \frac{1}{2} \mathbf{S}_\mu = 0$$

which gives the following MLE solution for the multivariate Gaussian's covariance matrix

$$\boxed{\boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^\top}$$

Question 4

[redacted]

$y = \{y_1, y_2, \ldots, y_N\}$

$P(y|\lambda) = \frac{\lambda^{y_n} e^{-\lambda}}{y!}$ (Poisson)

$P(\lambda) = \frac{\beta^\alpha}{\Gamma\alpha} \lambda^{\alpha-1} e^{-\beta\lambda}$ (Gamma)

(1) MLE and MAP:

$$\hat{\lambda}_{MLE} = \underset{\lambda}{\operatorname{argmax}} \sum_{n=1}^{N} \log P(y_n|\lambda) \quad (\text{i.i.d data})$$

$$\sum_{n=1}^{N} \log P(y_n|\lambda) = \left[\sum_{n=1}^{N} y_n \log\lambda\right] - N\lambda + \text{constant}$$

(terms that don't depend on λ)

Taking derivative and setting it to zero

$$\sum y_n \times \frac{1}{\lambda} - N = 0$$

$$\Rightarrow \boxed{\hat{\lambda}_{MLE} = \frac{\sum_{n=1}^{N} y_n}{N}}$$

MAP estimation will be almost identical with the extra $\log P(\lambda)$ term.

$$\hat{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}} \left[\sum_{n=1}^{N} \log P(y_n|\lambda) + \log P(\lambda)\right]$$

Scanned by CamScanner

__/__/____

$$\log p(\lambda) = (\alpha-1)\log\lambda - \beta\lambda + \text{Constant}$$

Constant: terms that don't depend on λ

The MAP objective will be

$$\sum_{n=1}^{N} y_n \log\lambda - N\lambda + (\alpha-1)\log\lambda - \beta\lambda$$

Maximizing w.r.t. λ will give the MAP solution

$$\hat{\lambda}_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N+\beta}$$

(2) Posterior distribution of λ

$$p(\lambda|y) = \frac{p(\lambda)p(y|\lambda)}{p(y)} = \frac{p(\lambda)\prod_{n=1}^{N} p(y_n|\lambda)}{p(y)}$$

Since the prior (gamma) and the likelihood (Poisson) are conjugate, the posterior must be gamma!

Let's multiply the terms $p(\lambda)$ and $p(y_n|\lambda)$ are try to "identify" this gamma distribution's parameters

$$p(\lambda|y) \propto p(y)\prod_{n=1}^{N} p(y_n|\lambda)$$

Scanned by CamScanner

$$P(\lambda|y) \propto \lambda^{\alpha-1} e^{-\beta\lambda} \times \prod_{n=1}^{N} \frac{\lambda^{y_n} e^{-\lambda}}{y_n!}$$

$$\propto \lambda^{\sum_{n=1}^{N} y_n + \alpha - 1} e^{-(\beta+N)\lambda}$$

(ignoring the terms that don't depend on λ)

The above expression is clearly in form of a gamma distribution with

$$\text{Shape} = \sum_{n=1}^{N} y_n + \alpha$$

$$\text{rate} = \beta + N$$

[No need to worry about constant of proportionality. It must be a gamma]

Thus $\boxed{P(y|\lambda) = \text{Gamma}\left(\sum_{n=1}^{N} y_n + \alpha, \beta + N\right)}$

(3) $$\hat{\lambda}_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N+\beta} = \frac{\sum_{n=1}^{N} y_n}{N+\beta} + \frac{(\alpha-1)}{(N+\beta)}$$

Sum to 1 (Convex comb.)

$$= \left[\sum_{n=1}^{N} \frac{y_n}{N}\right] \times \frac{N}{N+\beta} + \left[\frac{\alpha-1}{\beta}\right] \times \frac{\beta}{N+\beta}$$

↓ mLE ↓ Prior's mode

Posterior's mean:

$$\frac{\sum_{n=1}^{N} y_n + \alpha}{N+\beta} = \frac{\sum_{n=1}^{N} y_n}{N} \times \left[\frac{N}{N+\beta}\right] + \left[\frac{\alpha}{\beta}\right] \times \frac{\beta}{N+\beta}$$

↓ mLE ↓ Prior's mean

Scanned by CamScanner

(4) $P(y_* | y) = \int P(y_*, \lambda | y) d\lambda$

$= \int P(y_* | \lambda) P(\lambda | y) d\lambda$

$\approx P(y_* | \hat{\lambda}_{MLE})$ (if using MLE)

or

$P(y_* | \hat{\lambda}_{MAP})$ (if using MAP)

In both these cases, $P(y_* | y)$ is simply a Poisson with parameters λ_{MLE} or λ_{MAP}

If using the full posterior, which is $P(\lambda | y) = \text{Gamma}(\sum_{n=1}^{N} y_n + \alpha, \beta + N)$, $P(y_* | y)$ will be

$$P(y_* | y) = \int P(y_* | \lambda) P(\lambda | y) d\lambda$$

$$= \int \frac{\lambda^{y_*} e^{-\lambda}}{y_*!} \times \frac{(\beta+N)^{\sum_{n=1}^{N} y_n + \alpha}}{\left[\sum_{n=1}^{N} y_n + \alpha\right]} \lambda^{\sum_{n=1}^{N} y_n + \alpha - 1} e^{-(\beta+N)\lambda} d\lambda$$

The above is actually a mixture of infinite many Poisson distributions. The result is actually not a Poisson but the "Negative Binomial" distribution.

(if you are interested in knowing more about this result, you may refer to the wikipedia article of NB distribution).

(This part was just for your "General knowledge" :-))

Scanned by CamScanner

[redacted]

Question 5

$X = \{x_1, x_2, \ldots x_N\}$

$y = \{y_1, y_2, \ldots y_N\}$

The likelihood (both X and y are modeled here)

$$P(X, y \mid \theta) = \prod_{n=1}^{N} P(x_n, y_n \mid \theta)$$

(θ: All the params) $\theta = \{\mu_k, \Sigma_k, \pi_k\}_{k=1}^{K}$

Indicator function: $\mathbb{I}[y_n = k]$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \left[P(x_n, y_n = k \mid \theta) \right]$$

Note that only one term in the product will be selected based on the value of y_n

Thus log-likelihood will be

$$\log P(X, y \mid \theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \left[\log P(x_n, y_n = k \mid \theta) \right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \left[\log P(x_n \mid y_n = k, \theta) P(y_n = k \mid \theta) \right]$$

$P(x_n \mid y_n = k, \theta) \rightarrow \mathcal{N}(x_n \mid \mu_k, \Sigma_k)$; $P(y_n = k \mid \theta) \rightarrow \pi_k$

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}[y_n = k] \left[\log \mathcal{N}(x_n \mid \mu_k, \Sigma_k) + \log \pi_k \right]$$

To optimize w.r.t. π_k, we take partial derivatives w.r.t. π_k <u>BUT</u> need to use the constraint $\sum_{k=1}^{K} \pi_k = 1$. Also note that for $\{\pi_k\}_{k=1}^{K}$, the part of the objective that matters is

Scanned by CamScanner

$$\mathcal{L}(\pi_1, \ldots \pi_K) = \sum_{n=1}^{N} \sum_{K=1}^{K} \mathbb{I}[y_n = k] \log \pi_k$$

(Lagrangian)

Need to min. over π, max over λ

$$= \sum_{K=1}^{K} N_K \log \pi_K$$

N_K = # of points with $y_n = K$

↑ The Constrained formulation will be

$$\mathcal{L}(\pi, \lambda) = \sum_{K=1}^{K} N_K \log \pi_K + \lambda \left[\sum_{K=1}^{K} \pi_K - 1 \right]$$

Taking partial derivative w.r.t. π_K,

$$\frac{N_K}{\pi_K} + \lambda = 0$$

$$\pi_K = -\frac{N_K}{\lambda}$$

Also, using $\sum_{K=1}^{K} \pi_K = 1$ gives $\lambda = -N$

thus $\boxed{\pi_K = \frac{N_K}{N}}$

Now, for μ_K and Σ_K :

The relevant part of the objective function is

$$\mathcal{L}\left(\{\mu_K, \Sigma_K\}_{K=1}^{K}\right) = \sum_{n=1}^{N} \sum_{K=1}^{K} \mathbb{I}[y_n = k] \log \mathcal{N}(x_n | \mu_K, \Sigma_K)$$

Scanned by CamScanner

MLE for μ_k, Σ_k is exactly the same as the MLE procedure for a multivariate Gaussian, but only using inputs x_n from class k. [redacted])

(See attached slides for Gaussian MLE)

I have provided additional slides on the class webpage showing how to do it for multivariate Gaussian. The procedure for this problem will be exactly the same (but only using x_n with $y_n = k$)

Scanned by CamScanner

Gaussian Distribution

- The (multivariate) Gaussian with mean $\boldsymbol{\mu}$ and cov. matrix $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\Big\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big\} \\ &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\Big\{-\frac{1}{2}\text{trace}\Big[\boldsymbol{\Sigma}^{-1}\mathbf{S}\Big]\Big\} \qquad \text{where } \mathbf{S}=(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\end{aligned}$$

- An alternate representation: The "information form"

$$\mathcal{N}_c(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\Lambda}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2}\exp\Big\{-\frac{1}{2}\Big(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x}+\boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}-2\boldsymbol{x}^\top\boldsymbol{\xi}\Big)\Big\}$$

  where $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ are the "natural parameters" (recall exp. family).

- Note that there is a term quadratic in $\boldsymbol{x}$ (involves $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$) and linear in $\boldsymbol{x}$ (involves $\boldsymbol{\xi}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$)
- Information form can help recognize $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of a Gaussian when doing algebraic manipulations

Estimating Parameters of Gaussian: MLE

- Given: N i.i.d. observations $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ from a multivariate Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Goal: Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Simple to do MLE for this task

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Plugging in $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$ and ignoring the constants

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) \\
&= \frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\sum_{n=1}^{N}\text{trace}[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^\top] \\
&= \frac{N}{2}\log|\boldsymbol{\Sigma}|^{-1} - \frac{1}{2}\text{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}_\mu] \qquad \left[\text{where} \quad \mathbf{S}_\mu = \sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^\top\right]
\end{aligned}$$

Estimating Parameters of Gaussian: MLE

- Taking (partial) derivatives w.r.t. $\boldsymbol{\mu}$ and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\partial}{\partial \boldsymbol{\mu}} \left[\frac{N}{2} \log |\boldsymbol{\Sigma}|^{-1} - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \right] = -\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-\top})(\boldsymbol{x}_n - \boldsymbol{\mu}) = 0$$

which gives the following MLE solution for the multivariate Gaussian's mean

$$\boxed{\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n}$$

- Taking derivatives w.r.t. $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ (instead of $\boldsymbol{\Sigma}$; leads to simpler derivatives) and setting to zero

$$\frac{\partial}{\partial \boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{\partial}{\partial \boldsymbol{\Lambda}} \left[\frac{N}{2} \log |\boldsymbol{\Lambda}| - \frac{1}{2} \text{trace}[\boldsymbol{\Lambda} \mathbf{S}_\mu] \right] = \frac{N}{2} \boldsymbol{\Lambda}^{-\top} - \frac{1}{2} \mathbf{S}_\mu^\top = \frac{N}{2} \boldsymbol{\Lambda}^{-1} - \frac{1}{2} \mathbf{S}_\mu = \frac{N}{2} \boldsymbol{\Sigma} - \frac{1}{2} \mathbf{S}_\mu = 0$$

which gives the following MLE solution for the multivariate Gaussian's covariance matrix

$$\boxed{\boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^\top}$$

Question 6

Skipping the solution to practice problem 6 because we had already seen the Beta-Bernoulli case in class and the derivations for the Dirichlet-multinoulli case are almost identical.

QUESTION 7

Consider a logistic regression model $\mathsf{p}(y_n|\mathbf{x}_n, \mathbf{w}) = \frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}$, with a zero-mean Gaussian prior $\mathsf{p}(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$ and $y_n \in \{+1, -1\}$.

$\therefore \mathsf{p}(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}$ (Assuming i.i.d.)

For MAP estimation we have $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathsf{p}(\mathbf{w}|\mathbf{y}, \mathbf{X})$

$$\mathsf{p}(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{\mathsf{p}(\mathbf{y}|\mathbf{w}, \mathbf{X})\mathsf{p}(\mathbf{w})}{\mathsf{p}(\mathbf{y}|\mathbf{X})}$$

Since $\mathsf{p}(\mathbf{y}|\mathbf{X})$ is independent of $\mathbf{w}$ we have $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathsf{p}(\mathbf{y}|\mathbf{w}, \mathbf{X})\mathsf{p}(\mathbf{w})$

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} -\log(\mathsf{p}(\mathbf{y}|\mathbf{w}, \mathbf{X})) - \log(\mathsf{p}(\mathbf{w})) \\
&= \arg\min_{\mathbf{w}} -\log(\prod_{n=1}^{N} \frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} + constant \\
&= \arg\min_{\mathbf{w}} -\log(\prod_{n=1}^{N} \frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} \\
&= \arg\min_{\mathbf{w}} -\sum_{n=1}^{N} \log(\frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}
\end{aligned}$$

Consider $\mathcal{L}(\mathbf{w}) = -\sum_{n=1}^{N} \log(\frac{1}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}) + \frac{\lambda^2}{2}\mathbf{w}^T\mathbf{w}$

$$\frac{\partial\mathcal{L}(\mathbf{w})}{\partial\mathbf{w}} = -\sum_{n=1}^{N} \frac{\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)} y_n\mathbf{x}_n + \lambda\mathbf{w}$$

Setting $\frac{\partial\mathcal{L}(\mathbf{w})}{\partial\mathbf{w}} = 0$, we have

$$\begin{aligned}
\hat{\mathbf{w}} &= \frac{1}{\lambda}\sum_{n=1}^{N} \frac{\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)} y_n\mathbf{x}_n \\
&= \sum_{n=1}^{N} \alpha_n y_n\mathbf{x}_n
\end{aligned}$$

where $\alpha_n = \frac{1}{\lambda}\frac{\exp(-y_n\mathbf{w}^T\mathbf{x}_n)}{1+\exp(-y_n\mathbf{w}^T\mathbf{x}_n)} = \frac{1}{\lambda}\frac{1}{1+\exp(y_n\mathbf{w}^T\mathbf{x}_n)}$.

α_n can also be written as $\frac{1}{\lambda}(1 - \mathsf{p}(y_n|\mathbf{x}_n))$ which is nothing but the probability of getting incorrect label on input data $\mathbf{x}_n$ scaled by a factor of $\frac{1}{\lambda}$. Now consider if a small probability of getting correct label with current $\mathbf{w}$ the $\frac{1}{\lambda}(1 - \mathsf{p}(y_n|\mathbf{x}_n))$ or α_n will be large contribution from this $\mathbf{x}_n, y_n$ in calculating new $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n\mathbf{x}_n$ (taking a particular case of GD update when $\lambda = 1$). Now we attempt to show this change will improve $\mathbf{w}$ w.r.t this particular $\mathbf{x}_n$. This is very similar to a perceptron update (mistake-driven), where the input having incorrect prediction will lead to significant change in weight vector. Also a factor of $\frac{1}{\lambda}$ will act as a regularizer to regulate the contribution of each $\mathbf{x}_n$.

[redacted]

QUESTION 8

[redacted]

For the given problem class-marginal distribution $\mathsf{p}(y=1)=\pi$ and each class-conditional distribution is be defined as a product of D Bernoulli distributions, i.e., $\mathsf{p}(\mathbf{x}|y=1)=\prod_{d=1}^{D}\mathsf{p}(x_d|y=1)$ and $\mathsf{p}(\mathbf{x}|y=0)=\prod_{d=1}^{D}\mathsf{p}(x_d|y=0)$ where $\mathsf{p}(x_d|y=1)=Bernoulli(x_d|\mu_{d,1})$ and $\mathsf{p}(x_d|y=0)=Bernoulli(x_d|\mu_{d,0})$.
Assuming all the parameters have been estimated from the data, we can write the predictive distribution as follows

$$\begin{aligned}
\mathsf{p}(y=1|\mathbf{x}) &= \frac{\mathsf{p}(y=1)\mathsf{p}(\mathbf{x}|y=1)}{\mathsf{p}(y=1)\mathsf{p}(\mathbf{x}|y=1)+\mathsf{p}(y=0)\mathsf{p}(\mathbf{x}|y=0)} \\
&= \frac{\pi\prod_{d=1}^{D}\mathsf{p}(x_d|y=1)}{\pi\prod_{d=1}^{D}\mathsf{p}(x_d|y=1)+(1-\pi)\prod_{d=1}^{D}\mathsf{p}(x_d|y=0)} \\
&= \frac{1}{1+\frac{(1-\pi)\prod_{d=1}^{D}\mathsf{p}(x_d|y=0)}{\pi\prod_{d=1}^{D}\mathsf{p}(x_d|y=1)}} \\
&= \frac{1}{1+\frac{1-\pi}{\pi}\frac{\prod_{d=1}^{D}\mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}}{\prod_{d=1}^{D}\mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}}}
\end{aligned}$$

Comparing the above expression with $\sigma(\mathbf{w}^T\mathbf{x})$ i.e. $\exp(-(\mathbf{w}^T\mathbf{x}+b))=\frac{1-\pi}{\pi}\frac{\prod_{d=1}^{D}\mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d}}{\prod_{d=1}^{D}\mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}}$ we have

$$\begin{aligned}
-\mathbf{w}^T\mathbf{x}-b &= \log(\frac{1-\pi}{\pi})+\sum_{d=1}^{D}\log(\mu_{d,0}^{x_d}(1-\mu_{d,0})^{1-x_d})-\sum_{d=1}^{D}\log(\mu_{d,1}^{x_d}(1-\mu_{d,1})^{1-x_d}) \\
&= \log(\frac{1-\pi}{\pi})+\sum_{d=1}^{D}x_d\log\mu_{d,0}+\sum_{d=1}^{D}(1-x_d)\log(1-\mu_{d,0})-\sum_{d=1}^{D}x_d\log\mu_{d,1} \\
&\quad -\sum_{d=1}^{D}(1-x_d)\log(1-\mu_{d,1}) \\
&= \sum_{d=1}^{D}x_d\log(\frac{\mu_{d,0}(1-\mu_{d,1})}{(1-\mu_{d,0})\mu_{d,1}})+\log(\frac{1-\pi}{\pi}\prod_{d=1}^{D}\frac{1-\mu_{d,0}}{1-\mu_{d,1}})
\end{aligned}$$

From above we can see that we can find an equivalent probabilistic discriminative model for this problem with

$$\mathbf{w} = \begin{bmatrix} \log(\frac{\mu_{1,1}(1-\mu_{1,0})}{(1-\mu_{1,1})\mu_{1,0}}) \\ \log(\frac{\mu_{2,1}(1-\mu_{2,0})}{(1-\mu_{2,1})\mu_{2,0}}) \\ \vdots \\ \log(\frac{\mu_{D,1}(1-\mu_{D,0})}{(1-\mu_{D,1})\mu_{D,0}}) \end{bmatrix} \tag{1}$$

$$b = \log(\frac{\pi}{1-\pi}\prod_{d=1}^{D}\frac{1-\mu_{d,1}}{1-\mu_{d,0}}) \tag{2}$$

Having found $\mathbf{w}$ and b we can say that we have a linear decision boundary given by hyper-plane equation $\mathbf{w}^T\mathbf{x} + b = 0$