Learning with Gaussians

CPSC 440/550: Advanced Machine Learning

cs.ubc.ca/~dsuth/440/23w2

University of British Columbia, on unceded Musqueam land

2023-24 Winter Term 2 (Jan–Apr 2024)

Admin

admin

- Project guidelines finally available
- Brief proposal due March 29th
  - 10% of project grade, “lightly graded”: mostly checking scope of project
  - If you hand in earlier, we’ll give you scope feedback earlier
- Actual project due last day of finals (Saturday, April 27)
  - 6-page writeup, plus possible appendices/code supplement
  - Details on format to come

Last time: Multivariate Gaussians

- Continuous density estimation, $d > 1$ with the multivariate Gaussian distribution

$$x \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{means} \quad p(x \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}} \, |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}) \right)$$

- If $\boldsymbol{\Sigma}$ is a diagonal matrix, product of univariate normals
- $\boldsymbol{\mu}_j$ is $\mathbb{E}[x_j]$; $\boldsymbol{\Sigma}_{jj'}$ gives $\operatorname{Cov}(x_j, x_{j'})$
  - If $\operatorname{Cov}(x_j, x_{j'}) = 0$, then $x_j \perp\!\!\!\perp x_{j'}$ (for jointly-Gaussian variables)
- If $\boldsymbol{\Sigma}$ is singular, "degenerate" Gaussian: $v^{\mathsf{T}} x$ takes a constant value for some v
- $Ax + b \sim \mathcal{N}(A\boldsymbol{\mu} + b, A\boldsymbol{\Sigma}A^{\mathsf{T}}$
  - Lets us sample based on $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - Marginalizing: still normal, just ignore the other variables in $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$
  - Conditioning: $x \mid z \sim \mathcal{N}\left(\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xz}\boldsymbol{\Sigma}_z^{-1}(z - \boldsymbol{\mu}_z), \boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_{xz}\boldsymbol{\Sigma}_z^{-1}\boldsymbol{\Sigma}_{xz}^{\mathsf{T}}\right)$

Outline

MLE for the mean of a multivariate Gaussian

- If $x^{(i)} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $\boldsymbol{\Sigma} \succ 0$, we have

$$p\left(x^{(i)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(x^{(i)} - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}\left(x^{(i)} - \boldsymbol{\mu}\right)\right),$$

  so up to a constant our negative log-likelihood for n examples is

$$\frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)} - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}\left(x^{(i)} - \boldsymbol{\mu}\right) + \frac{n}{2}\log|\boldsymbol{\Sigma}|$$

- This is a convex quadratic in μ; setting gradient to zero gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

  - Mean along each dimension; it doesn't depend on Σ

MLE for the covariance of a multivariate Gaussian

- To get MLE for Σ we can re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$\frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right) + \frac{n}{2}\log|\boldsymbol{\Sigma}|$$

$$= \frac{1}{2}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}} \boldsymbol{\Theta}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right) + \frac{n}{2}\log\left|\boldsymbol{\Theta}^{-1}\right|$$

- After some work (bonus slides), we get that this is equal to

$$f(\boldsymbol{\Theta}) = \frac{n}{2}\mathrm{Tr}(\mathbf{S}\Theta) - \frac{n}{2}\log|\boldsymbol{\Theta}|, \text{ with } \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}}$$

  - $\mathbf{S}$ is the sample covariance: if $\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n\hat{\mu}^{\mathsf{T}}$ is centred data, $S = \frac{1}{n}\tilde{\mathbf{X}}^{\mathsf{T}}\tilde{\mathbf{X}}$
  - Trace operator $\mathrm{Tr}(A)$ is the sum of the diagonal elements of A
  - $\mathrm{Tr}(A^{\mathsf{T}}B) = \sum_j (A^{\mathsf{T}}B)_{jj} = \sum_j \sum_i (A^{\mathsf{T}})_{ji}B_{ij} = \sum_{ij} A_{ij}B_{ij}$, i.e. `(A * B).sum()`

MLE for the covariance of a multivariate Gaussian

- Gradient matrix of NLL with respect to Θ is (not obvious, see bonus slides)

$$\nabla f(\Theta) = \frac{n}{2}\left(\mathbf{S} - \mathbf{\Theta}^{-1}\right) \quad \text{for } S = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}}$$

- The MLE for a given μ is obtained by setting the gradient matrix to zero, giving

$$\Theta = \mathbf{S}^{-1} \quad \text{or} \quad \Sigma = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}}$$

- To have $\Sigma \succ 0$, we need a positive-definite sample covariance, $S \succ 0$
  - If S is not positive definite, NLL is unbounded below, and MLE doesn't exist
  - Like requiring "not all values are the same" in univariate Gaussian
  - In d dimensions, you need d linearly independent $x^{(i)}$ values (no "multi-collinearity")
  - This is only possible if $n \geq d$! (But might not be true even if it is)

- Note: most distributions' MLEs don't correspond with "moment matching"

Example: Multivariate Gaussians on MNIST

- Let's try continuous density estimation on (binary) handwritten digits

$x^i = \mathrm{vec}\Big($  $\Big) \in \mathbb{R}^{784}$

7 7

4 4

Diagonal Σ:

$\hat{\mu} = \mathrm{vec}\Big($  $\Big)$

$\hat{\Sigma} = \mathrm{diag}\Big(\mathrm{vec}\Big($  $\Big)\Big)$

General Σ:

$\hat{\mu}$ is the same (!)

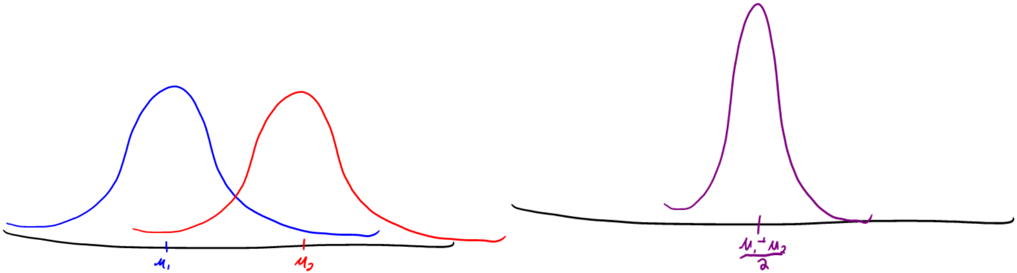
$\hat{\Sigma}$ is big (784 by 784)

Product of Gaussian densities

- This property will be helpful in deriving MAP/Bayesian estimation:
- Consider a variable x whose pdf is written as product of two Gaussians,

$$p(x) \propto \underbrace{\mathcal{N}(x \mid \boldsymbol{\mu}_1, \mathbf{I})}_{\text{density of } \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{I}) \text{ at } x} \quad \mathcal{N}(x \mid \boldsymbol{\mu}_2, \mathbf{I})$$

- This product of Gaussian pdfs is a Gaussian with $\boldsymbol{\mu} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ and $\boldsymbol{\Sigma} = \frac{1}{2}\mathbf{I}$



Product of Gaussian densities

- If $p(x) \propto \mathcal{N}(x \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\,\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$,
  then x is Gaussian with (see PML2 2.2.7.6 – complete the square in the exponent)

$$\text{covariance} \quad \boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})^{-1}$$

$$\text{mean} \quad \boldsymbol{\mu} = \boldsymbol{\Sigma}\,\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

- Consider $x^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for fixed $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$:

$$\begin{aligned}
p(\boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &\propto p(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} p\left(x^{(i)} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) && \text{(Bayes rule)} \\
&= p(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} p(\boldsymbol{\mu} \mid x^{(i)}, \boldsymbol{\Sigma}) && \text{(symmetry of } x^{(i)} \text{ and } \boldsymbol{\mu}) \\
&= (\text{product of } (n+1) \text{ Gaussians})
\end{aligned}$$

- So, working it out gives. . .

MAP estimation for mean

- For fixed Σ, conjugate prior for mean is a Gaussian:

$$x^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \mu \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad \text{implies} \quad \boldsymbol{\mu} \mid \mathbf{X}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma}^+),$$

  where

$$\boldsymbol{\Sigma}^+ = (n\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1},$$
$$\boldsymbol{\mu}^+ = \Sigma^+(n\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\text{MLE}} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \qquad \text{MAP estimate of } \mu$$

- In special case of $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ and $\Sigma_0 = \frac{1}{\lambda}\mathbf{I}$, we get

$$\boldsymbol{\Sigma}^+ = \left(\frac{n}{\sigma^2}\mathbf{I} + \lambda\mathbf{I}\right)^{-1} = \frac{1}{\frac{n}{\sigma^2} + \lambda}\mathbf{I},$$
$$\boldsymbol{\mu}^+ = \boldsymbol{\Sigma}^+\left(\frac{n}{\sigma^2}\boldsymbol{\mu}_{\text{MLE}} + \lambda\boldsymbol{\mu}_0\right)$$

- Posterior predictive is $\mathcal{N}(\boldsymbol{\mu}^+, \boldsymbol{\Sigma} + \boldsymbol{\Sigma}^+)$ – take product of $(n+2)$ then marginalize
  - Many Bayesian inference tasks have closed form; if not, Monte Carlo is easy

MAP Estimation in Multivariate Gaussian (Trace Regularization)

- A common MAP estimate for Σ is

$$\hat{\Sigma} = \mathbf{S} + \lambda \mathbf{I},$$

where S is the covariance of the data
  - Key advantage: $\hat{\Sigma}$ is strictly positive definite (eigenvalues are at least λ)

- This corresponds to L1 regularization of precision diagonals (see bonus)

$$f(\Theta) = \underbrace{\mathrm{Tr}(\mathbf{S}\mathbf{\Theta}) - \log|\mathbf{\Theta}|}_{\text{NLL times } 2/n} + \lambda \sum_{j=1}^{d} |\mathbf{\Theta}_{jj}|$$

- Note this *doesn't* set Θ_{jj} values to exactly zero
  - Log-determinant term becomes arbitrarily steep as the Θ_{jj} approach 0
  - It's not really the case that "L1 gives sparsity"; it's "L2 + L1 gives sparsity"

Conjugate Priors for Covariance

bonus!

- Trace regularization (or Graphical LASSO, later): not a conjugate prior

- Conjugate prior for Θ with known mean is Wishart distribution
  - A multi-dimensional generalization of the gamma distribution
    - Gamma is a distribution over positive scalars
    - Wishart is a distribution over positive-definite matrices
  - Posterior predictive is a student t distribution
  - Conjugate prior for Σ is inverse-Wishart (equivalent posterior)

- If both μ and Θ are variables, conjugate prior is normal-Wishart
  - Normal times Wishart, with a particular dependency among parameters
  - Posterior predictive is again a student t distribution

- Wikipedia has already done a lot of possible homework questions for you:
  - `https://en.wikipedia.org/wiki/Conjugate_prior`

Outline

Generative Classification with Gaussians

- Consider a generative classifier with continuous features:

$$p(y \mid x) \propto p(x, y) = \underbrace{p(x \mid y)}_{\text{continuous}} \underbrace{p(y)}_{\text{discrete}}$$

- Model y as a categorical distribution (classification task)
- Previously handled $p(x \mid y)$ with the naive Bayes assumption, $x_i \perp\!\!\!\perp x_j \mid y$
  - Strong, usually unrealistic assumption

- In Gaussian discriminant analysis (GDA) we assume $x \mid y$ is Gaussian
  - Classifier asks "which Gaussian makes this $x^{(i)}$ most likely?"
  - This can model pairwise correlations within each class
    - Doesn't need the naive Bayes assumption

Gaussian Discriminant Analysis (GDA)

- In Gaussian discriminant analysis we assume $x \mid y$ is Gaussian

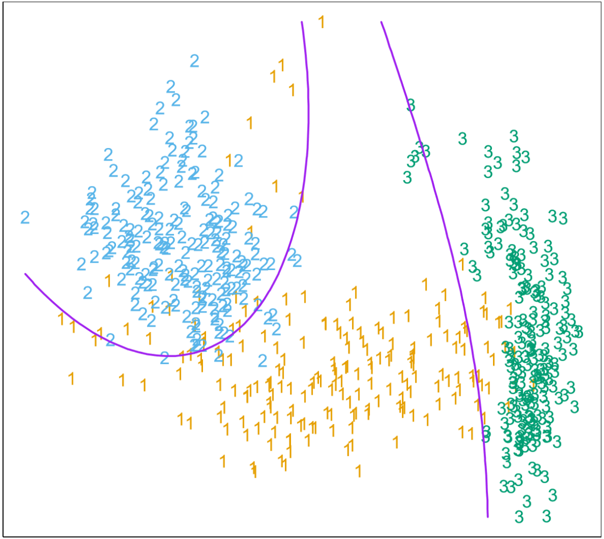
$$p(x, y = c) = \underbrace{p(y)\, p(x \mid y = c)}_{\text{product rule}} = \underbrace{\pi_c}_{\Pr(y=c)} \underbrace{p(x \mid \boldsymbol{\mu_c}, \boldsymbol{\Sigma_c})}_{\text{Gaussian pdf}}$$

- Classify based on

$$\begin{aligned} \arg\max_c p(y = c \mid x) &= \arg\max_c \log p(y = c, x) \\ &= \arg\max_c \ \log \pi_c - \frac{1}{2} \log |\boldsymbol{\Sigma_c}| - \frac{1}{2} (x - \boldsymbol{\mu_c})^{\mathsf{T}} \boldsymbol{\Sigma_c}^{-1} (x - \boldsymbol{\mu_c}) \end{aligned}$$

- With general choices for $\boldsymbol{\mu_c}$ and $\boldsymbol{\Sigma_c}$, we're taking the max of k quadratics
  - Means that the decision boundary will be zeros of a quadratic ("quadric surface")
  - Leads to the equivalent name quadratic discriminant analysis (QDA)
- Fitting GDA=QDA: fit π_c as categorical, fit Gaussian for each subset with $y^{(i)} = c$

GDA=QDA example



https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Special case: Linear Discriminant Analysis (LDA)

- A common special case: constrain $\boldsymbol{\Sigma}_{\boldsymbol{c}} = \boldsymbol{\Sigma}$ for all c
- Means that we classify as

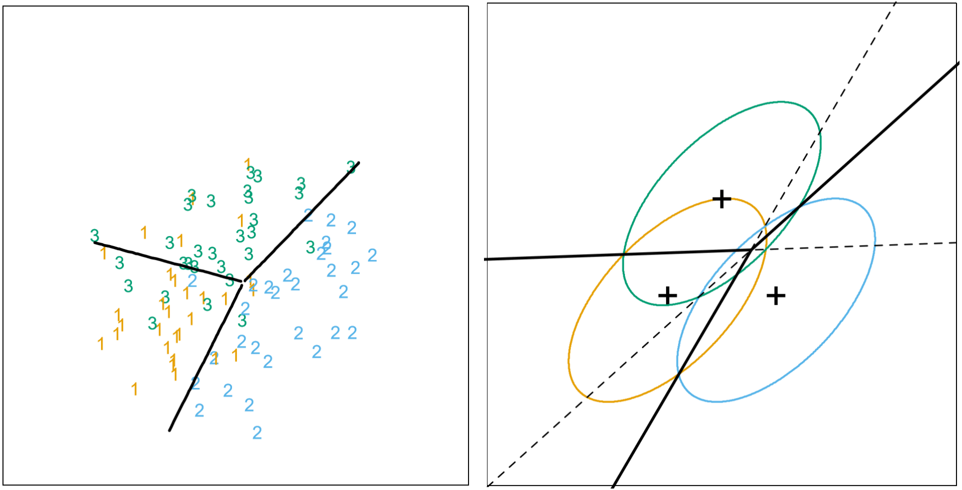
$$\begin{aligned}
\arg\max_c p(y = c \mid x) &= \arg\max_c \; \log \pi_c - \tfrac{1}{2} \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} (x - \boldsymbol{\mu}_{\boldsymbol{c}})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}_{\boldsymbol{c}}) \\
&= \arg\max_c \; \log \pi_c - \tfrac{1}{2} x^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} x + \boldsymbol{\mu}_{\boldsymbol{c}}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} x - \tfrac{1}{2} \boldsymbol{\mu}_{\boldsymbol{c}}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\boldsymbol{c}} \\
&= \arg\max_c \; \underbrace{(\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\boldsymbol{c}})}_{w_c}{}^{\mathsf{T}} x + \underbrace{\log \pi_c - \tfrac{1}{2} \boldsymbol{\mu}_{\boldsymbol{c}}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\boldsymbol{c}}}_{b_c}
\end{aligned}$$

  so this is a linear classifier!
  - Behaves (asymptotically) optimally if the assumptions are true: $x \mid y \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$
  - May be terrible if these assumptions aren't true
- MLE in this model is simple: $\boldsymbol{\mu}_{\boldsymbol{c}}$ is mean of the points with $y^{(i)} = c$,

  $\boldsymbol{\Sigma}$ is $\displaystyle\frac{1}{n} \sum_{i=1}^{n} \left(x^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right) \left(x^{(i)} - \boldsymbol{\mu}_{y^{(i)}}\right)^{\mathsf{T}}$

LDA example

- Example of fitting linear discriminant analysis (LDA) to a 3-class problem:



https://web.stanford.edu/~hastie/Papers/ESLII.pdf

LDA and nearest neighbour

bonus!

- We classify according to

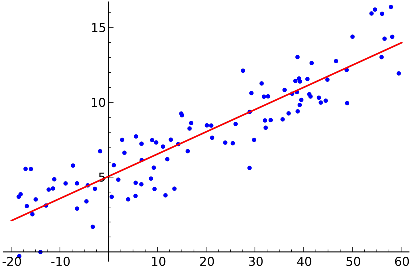
$$\begin{aligned}
&\arg\max_c \, (\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_c)^\mathsf{T}(\boldsymbol{\Sigma}^{-\frac{1}{2}}x) - \tfrac{1}{2}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_c)^\mathsf{T}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_c) + \log \pi_c \\
&\qquad = \arg\max_c \; -\tfrac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}x\|^2 + (\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_c)^\mathsf{T}(\boldsymbol{\Sigma}^{-\frac{1}{2}}x) - \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_c\|^2 + \log \pi_c \\
&\qquad = \arg\min_c \, \|\boldsymbol{\Sigma}^{-\frac{1}{2}}(x - \boldsymbol{\mu}_c)\|^2 - 2\log \pi_c
\end{aligned}$$

- If π_c are constant (all $\frac{1}{k}$) and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$, this picks the closest class mean
- With constant π_c but general $\boldsymbol{\Sigma}$, picks closest class mean in Mahalanobis distance

Outline

Regression with Gaussians

- In regression, y is continuous



https://en.wikipedia.org/wiki/Regression_analysis

- It's possible to use generative regression models (bonus slide)
  - For example, we could model $p(x, y)$ as a multivariate Gaussian
    - Then use that the conditional $p(y \mid x)$ is Gaussian for prediction

- But we usually treat features as fixed (as in discriminative classification models)
- Now ready to return to Bayesian linear regression

Bayesian Linear Regression

- Linear regression with Gaussian likelihood and prior,

$$y \mid x \sim \mathcal{N}(w^\top x, \sigma^2), \quad w \sim \mathcal{N}(0, \lambda^{-1}\mathbf{I})$$

- MAP estimate is ridge regression (L2-regularized least squares)
- Can use Gaussian identities to work out that the posterior has the form

$$w \mid (\mathbf{X}, \mathbf{y}) \sim \mathcal{N}\left(w_{\mathsf{MAP}}, \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\right),$$

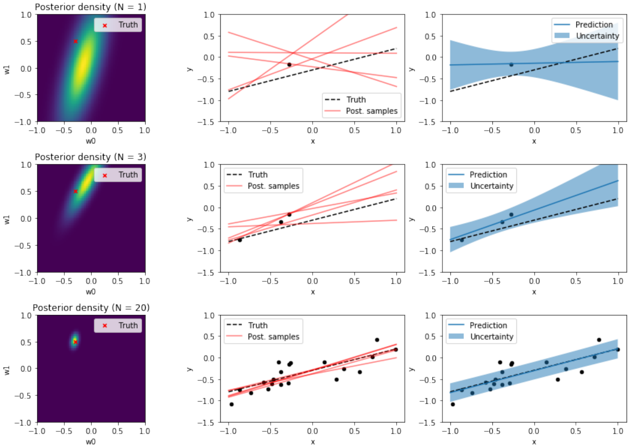
  which is a multivariate Gaussian centred at $w_{\mathsf{MAP}} = \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda}{\sigma^2}\mathbf{I}_d\right)^{-1}\mathbf{X}^\top\mathbf{y}$
  - The variance tells us how much variation we have around the MAP estimate
    - In other models, the posterior mode (MAP) is usually not the posterior mean
- By more Gaussian identities, the posterior predictive has the form

$$\tilde{y} \mid (\mathbf{X}, \mathbf{y}, \tilde{x}) \sim \mathcal{N}\left(w_{\mathsf{MAP}}^\top \tilde{x},\ \sigma^2 + \tilde{x}^\top\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\tilde{x}\right)$$

- Posterior predictive mode=mean again the MAP prediction in this model
  - Working with the full posterior predictive gives us variance of predictions

Bayesian Linear Regression

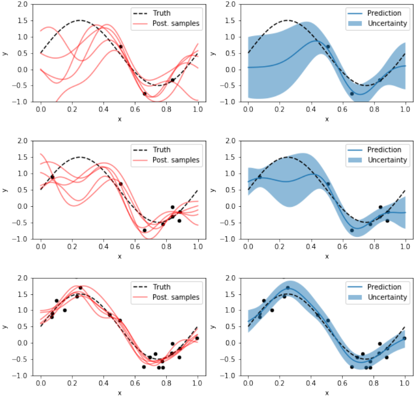
- Bayesian perspective gives us variability in w and predictions:



http://krasserm.github.io/2019/02/23/bayesian-linear-regression

Bayesian Linear Regression

- Bayesian linear regression with Gaussian RBFs as features:



http://krasserm.github.io/2019/02/23/bayesian-linear-regression

- We have not only a prediction, but Bayesian inference gives "error bars"
  - Gives an idea of "where model is confident" and where it is not

Digression: Gaussian Processes

bonus!

- In CPSC 340 you saw the kernel trick:
  - Rewrites L2-regularized least squares linear/prediction in terms of inner products
  - Allows us to efficiently use some exponential-sized or infinite-sized feature sets

- We can use kernel trick on posterior in Gaussian likelihood/prior model
  - Allows us to efficiently use some large or infinite-sized feature sets
  - Posterior in this case can be written as a Gaussian process (GP)

- Notation: a stochastic process is an infinite collection of random variables
- In a Gaussian process, any finite subcollection is jointly Gaussian
  - Defined in terms of a mean function and a covariance function
    - The set of possible covariance functions is the set of possible kernel functions
  - A popular book on this topic if you want to read more:
    Rasmussen/Williams, Gaussian Processes for Machine Learning

- We'll assume we have explicit features, but you could use kernels/GPs instead

Summary

- Gaussian discriminant analysis and special case linear discriminant analysis
  - Generative classifier where $x \mid y$ is multivariate normal
- Bayesian Linear Regression
  - Gaussian conditional likelihood and Gaussian prior gives Gaussian posterior
  - Posterior predictive is also Gaussian ("regression with error bars")

- Next time: choosing priors, sampling from complex posteriors

MLE for the covariance of a multivariate Gaussian

bonus!

- To get MLE for Σ we re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$\frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^\mathsf{T}\Sigma^{-1}(x^i - \mu) + \frac{n}{2}\log|\Sigma|$$

$$=\frac{1}{2}\sum_{i=1}^{n}(x^{(i)} - \mu)^\mathsf{T}\Theta(x^i - \mu) + \frac{n}{2}\log|\Theta^{-1}| \qquad \text{(okay because } \Sigma \text{ is invertible)}$$

$$=\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left((x^{(i)} - \mu)^\mathsf{T}\Theta(x^i - \mu)\right) + \frac{n}{2}\log|\Theta|^{-1} \qquad \text{(scalar } y^\mathsf{T}Ay = \mathrm{Tr}(y^\mathsf{T}Ay))$$

$$=\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}((x^{(i)} - \mu)(x^i - \mu)^\mathsf{T}\Theta) - \frac{n}{2}\log|\Theta| \qquad (\mathrm{Tr}(ABC) = \mathrm{Tr}(CAB))$$

- $\left|A^{-1}\right| = 1/\left|A\right|$ (can see e.g. from eigenvalues)
- The trace is the sum of the diagonal elements: $\mathrm{Tr}(A) = \sum_i A_{ii}$
  - $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ when dimensions match: called trace rotation or cyclic property

MLE for the covariance of a multivariate Gaussian

bonus!

- From the last slide,

$$p(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{Tr}\left(\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}} \boldsymbol{\Theta}\right) - \frac{n}{2}\log|\boldsymbol{\Theta}|$$

- We can exchange the sum and trace (trace is a linear operator) to get,

$$= \frac{1}{2}\,\mathrm{Tr}\left(\sum_{i=1}^{n}(x^{(i)} - \hat{\boldsymbol{\mu}})(x^{i} - \hat{\boldsymbol{\mu}})^{\mathsf{T}}\Theta\right) - \frac{n}{2}\log|\Theta| \qquad \sum_i \mathrm{Tr}(A_i B) = \mathrm{Tr}\left(\sum_i A_i B\right)$$

$$= \frac{n}{2}\,\mathrm{Tr}\left(\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x^{i} - \hat{\boldsymbol{\mu}})(x^{i} - \hat{\boldsymbol{\mu}})^{\mathsf{T}}}_{\text{sample covariance, } S}\right)\Theta\right) - \frac{n}{2}\log|\Theta| \qquad \left(\sum_i A_i B\right) = \left(\sum_i A_i\right) B$$

MLE for the covariance of a multivariate Gaussian

bonus!

- So the NLL in terms of the precision matrix Θ and sample covariance S is

$$f(\Theta) = \frac{n}{2}\text{Tr}(S\Theta) - \frac{n}{2}\log|\Theta|, \text{ with } S = \frac{1}{n}\sum_{i=1}^{n}\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)\left(x^{(i)} - \hat{\boldsymbol{\mu}}\right)^{\mathsf{T}}$$

- Weird-looking but has nice properties:
  - $\text{Tr}(S\Theta)$ is linear function of Θ, with $\nabla_\Theta \text{Tr}(S\Theta) = S$

    (it's the matrix version of an inner product $s^{\mathsf{T}}\theta$; called "Frobenius inner product")
  - Negative log-determinant is strictly convex, and $\nabla_\Theta \log|\Theta| = \Theta^{-1}$

    (generalizes $\nabla \log|x| = 1/x$ for for $x > 0$)
- Using these two properties the gradient matrix has a simple form:

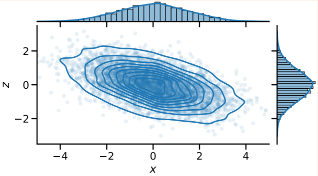
$$\nabla f(\Theta) = \frac{n}{2}(S - \Theta^{-1})$$

which is what we use to get the MLE

Generative Regression

bonus!

- With continuous features, we could model $p(x, y)$ as a multivariate Gaussian



- Training could use the closed-form MLE/MAP for multivariate Gaussian
- We obtain a univariate Gaussian $p(y \mid x)$ using conditioning formula,

$$y \mid x \sim \mathcal{N}\left(\mu_y + \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}(x - \boldsymbol{\mu}_x), \sigma_y^2 - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{yx}^{\mathsf{T}}\right)$$

- The conditional mean is a linear function, $w^{\mathsf{T}}x + b$
- Could extend to multiple outputs, with correlations given based on $\mathbf{\Sigma}_y$