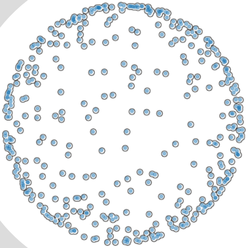
WiSe 2024/25

Machine Learning for Data Science

Lecture by G. Montavon

FREIE UNIVERSITÄT BERLIN



Lecture 9a | **Probabilistic Models**

Outline

Parameter Estimation

- Maximum Likelihood
- Bayesian Approach
- Weather Example

Estimating the Parameters of a Gaussian Distribution

- Estimating the Mean
- Estimating the Covariance

Probabilistic Inference

- Linear Regression Revisited
- Discriminant Revisited

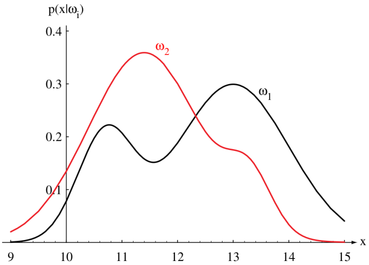
Classical vs. Probabilistic Approach

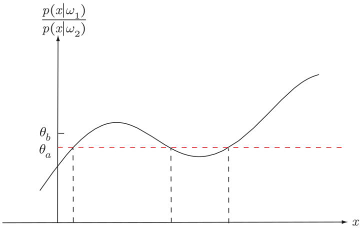
'Classical' approach:

- Define some statistic (e.g. variance in projected space) and search for the projection that maximizes it.

Probabilistic approach:

- **Step 1:** Learn a probability model of the data (e.g. assume the data comes from a Gaussian distribution and estimate its parameters).
- **Step 2:** Make predictions/inferences assuming the probability distributions and their parameters are the ground-truth.





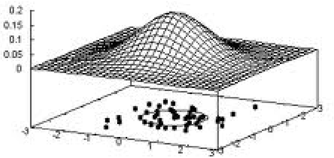
What is a Probability Model

A probability model consists of a probability law (assumed to be fixed) and its parameters (learned from the data).

Example 1:

- The multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ which returns for each point $\boldsymbol{x}$ the probability density

$$p(\boldsymbol{x} \,|\, \theta) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$$



What is a Probability Model

Example 2:

- The probability over a discrete set of possible observations $S_1, \dots, S_K$:

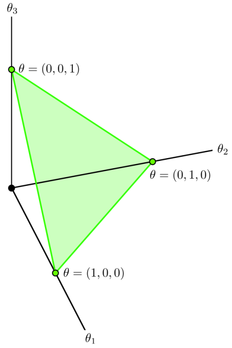
$$P(x \,|\, \theta) = \begin{cases} \theta_1 & \text{if } \; x = S_1 \\ \theta_2 & \text{if } \; x = S_2 \\ \vdots & \\ \theta_K & \text{if } \; x = S_K \end{cases}$$

with constraints

$$\theta_1, \theta_2, \dots, \theta_K \geq 0$$

and

$$\theta_1 + \theta_2 + \dots + \theta_K = 1$$



The Likelihood Function

- Assume that we have a dataset $\mathcal{D} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$.
- We now consider that $\boldsymbol{x}_i \in \mathbb{R}^d$ have been generated by our probability model (with density function $p(\boldsymbol{x} \,|\, \theta)$).
- If we further assume that all examples have been generated independently and identically distributed (iid.) from that distribution, we can express the probability density associated to our dataset $\mathcal{D}$ as:

$$\boxed{p(\mathcal{D} \,|\, \theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \,|\, \theta)}$$

- We call such function that depends on θ the *likelihood function*.

Two Approaches to Parameter Estimation:

Approach 1: Maximum Likelihood

- Find the parameter that is the most likely, i.e. for which the probability of the given dataset having been generated is the highest.

$$\theta^\star = \arg\max_\theta \overbrace{p(\mathcal{D} \,|\, \theta)}^{\text{likelihood function}}$$

Approach 2: Bayes

- Assume some initial distribution of parameters $\overbrace{p(\theta)}^{\text{prior distribution}}$, and refine this distribution in the light of the data, using the Bayes rule:

$$\Theta_\mathcal{D} \sim \overbrace{p(\theta \,|\, \mathcal{D})}^{\text{posterior distribution}} = \frac{\overbrace{p(\mathcal{D} \,|\, \theta)}^{\text{likelihood}}\,\overbrace{p(\theta)}^{\text{prior}}}{\int p(\mathcal{D} \,|\, \theta)\, p(\theta)\, d\theta}$$

Part 1

Maximum Likelihood

Maximum Likelihood: Weather Example

Assume whether observations are of the following type: 'sunny', 'cloudy', 'rainy'.

Let us define the following simple probability model of weather:

$$P(x \mid \theta) = \begin{cases} \alpha & \text{if} \quad x = \text{'sunny'} \\ \beta & \text{if} \quad x = \text{'cloudy'} \\ \gamma & \text{if} \quad x = \text{'rainy'} \end{cases}$$

where $\theta = (\alpha, \beta, \gamma)$ denotes the collection of parameters of our model. The parameters are subject to the constraints:

$$\alpha, \beta, \gamma \geq 0$$

and

$$\alpha + \beta + \gamma = 1.$$

Maximum Likelihood: Weather Example

Suppose we observe the following sequence of events (x_1, x_2, x_3, x_4):

sunny	cloudy	rainy	sunny

Making the assumption that the events have been generated iid. by our model, the likelihood function is given by

$$\begin{aligned} P(\mathcal{D} \,|\, \theta) &= \prod_{i=1}^{N} P(x_i \,|\, \theta) \\ &= \alpha \cdot \beta \cdot \gamma \cdot \alpha \\ &= \alpha^2 \cdot \beta \cdot \gamma \end{aligned}$$

and for practical purposes, we can also compute the log-likelihood:

$$\log P(\mathcal{D} \,|\, \theta) = 2 \log \alpha + \log \beta + \log \gamma$$

Maximum Likelihood: Weather Example

To find the parameters of the model that best explain the data, we can state the optimization problem:

$$\arg\max_{\theta} \left\{ 2\log\alpha + \log\beta + \log\gamma \right\} \quad \text{s.t.} \quad \alpha + \beta + \gamma = 1$$

We use the method of Lagrange multipliers, by first stating a Lagrange function:

$$\mathcal{L}(\theta;\lambda) = 2\log\alpha + \log\beta + \log\gamma + \lambda \cdot (1 - \alpha + \beta + \gamma)$$

and then finding points where the gradient of $\mathcal{L}$ is zero:

$$\partial\mathcal{L}/\partial\alpha = 2/\alpha - \lambda \overset{\text{(def)}}{=} 0 \quad \Rightarrow \quad \alpha = 2/\lambda$$

$$\partial\mathcal{L}/\partial\beta = 1/\beta - \lambda \overset{\text{(def)}}{=} 0 \quad \Rightarrow \quad \beta = 1/\lambda$$

$$\partial\mathcal{L}/\partial\gamma = 1/\gamma - \lambda \overset{\text{(def)}}{=} 0 \quad \Rightarrow \quad \gamma = 1/\lambda$$

Using the constraint $\alpha + \beta + \gamma = 1$ to eliminate the parameter λ, we get the maximum likelihood solution:

$$\alpha = \frac{1}{2}\,, \quad \beta = \frac{1}{4}\,, \quad \gamma = \frac{1}{4}\,.$$

Maximum Likelihood: Weather Example

So far, we have built a model of weather from four observations (x_1, x_2, x_3, x_4). Now, we would like to use it to predict future (unobserved) events.

Example:

- *Question:*
  What is the probability of next two events (x_5, x_6) being:

rainy	rainy

  Answer:

$$P(x_5 = \text{'rainy'} \,|\, \theta^\star) \cdot P(x_6 = \text{'rainy'} \,|\, \theta^\star) = \gamma \cdot \gamma = \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{16}$$

Part 2

Bayes Parameter Estimation

Bayes Parameter Estimation

Idea:

- Think of the infered parameter as a random variable $\Theta_{\mathcal{D}}$ following a distribution $p(\theta \,|\, \mathcal{D})$. The latter represent some prior distribution $p(\theta)$ refined in the light of the observations $\mathcal{D}$, and which can be obtained using the Bayes rule:

$$\Theta_{\mathcal{D}} \sim \overbrace{p(\theta \,|\, \mathcal{D})}^{\text{posterior distribution}} = \frac{\overbrace{p(\mathcal{D} \,|\, \theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\int p(\mathcal{D} \,|\, \theta)\, p(\theta)\, d\theta}$$

- Measuring likelihood of new data points $\mathcal{D}$ by integrating over all distributions of parameters.

$$\mathbb{E}[p(\mathcal{D}^\star \,|\, \Theta_{\mathcal{D}})] = \int \underbrace{p(\mathcal{D}^\star \,|\, \theta)}_{\text{likelihood}^\star} \underbrace{p(\theta \,|\, \mathcal{D})}_{\text{posterior}} d\theta$$

Bayes: Weather Example

Recall our weather example, where the three possible states are ‘sunny’, ‘cloudy’, and ‘rainy’.

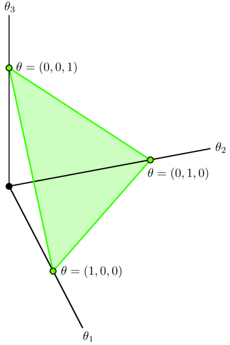
Step 1: Define the prior distribution

- Assuming we have no a priori knowledge, we can encode this lack of knowledge by a uniform prior distribution over the domain of parameters.

Such uniform distribution over the domain can be expressed by first noting that there are only two effective parameters (α, β), building the uniform distribution over these parameters:

$$p(\alpha, \beta) = \begin{cases} 2 & \alpha \in [0,1], \beta \in [0, 1-\alpha] \\ 0 & \text{else} \end{cases}$$

and recovering the parameter γ from the other two parameters as $\gamma = 1 - \alpha - \beta$.



Bayes: Weather Example

Recall that we made the following observations (x_1, x_2, x_3, x_4):

sunny	cloudy	rainy	sunny

Step 2: State the likelihood function

- We proceed similarly as in the maximum likelihood case:

$$\begin{aligned} P(\mathcal{D} \,|\, \theta) &= \prod_{i=1}^{N} P(x_i \,|\, \theta) \\ &= \alpha \cdot \beta \cdot \gamma \cdot \alpha \\ &= \alpha^2 \cdot \beta \cdot \gamma \end{aligned}$$

  and like for the prior distribution, express γ as a function of α and β:

$$\begin{aligned} P(\mathcal{D} \,|\, \theta) = P(\mathcal{D} \,|\, \alpha, \beta) &= \alpha^2 \cdot \beta \cdot (1 - \alpha - \beta) \\ &= \alpha^2 \beta - \alpha^3 \beta - \alpha^2 \beta^2 \end{aligned}$$

Bayes: Weather Example

Step 3: Compute the posterior distribution

- We get the posterior distribution by applying the Bayes rule and solving the integral:

$$
\begin{aligned}
p(\alpha,\beta \,|\, \mathcal{D}) &= \frac{\overbrace{P(\mathcal{D} \,|\, \alpha,\beta)}^{\text{likelihood}}\,\overbrace{p(\alpha,\beta)}^{\text{prior}}}{\int P(\mathcal{D} \,|\, \alpha,\beta)\, p(\alpha,\beta)\, d\alpha d\beta} \\
&= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)\cdot 2}{\int_0^1 \big(\int_0^{1-\alpha} (\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)\cdot 2 \cdot d\beta\big) d\alpha} \\
&= \dots \\
&= 360 \cdot (\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)
\end{aligned}
$$

... if $\alpha \in [0,1], \beta \in [0, 1-\alpha]$, else, $p(\alpha,\beta \,|\, \mathcal{D}) = 0$.

Bayes: Weather Example

Details of Step 3:

$$
\begin{aligned}
p(\alpha, \beta \,|\, \mathcal{D}) &= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)\cdot 2}{\int_0^1 \big(\int_0^{1-\alpha} (\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)\cdot 2 \cdot d\beta\big) d\alpha} \\
&= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)}{\int_0^1 \big((\alpha^2 \frac{\beta^2}{2} - \alpha^3 \frac{\beta^2}{2} - \alpha^2 \frac{\beta^3}{3})\big|_{\beta=0}^{1-\alpha}\big) d\alpha} \\
&= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)}{\int_0^1 \big(\alpha^2 \frac{(1-\alpha)^2}{2} - \alpha^3 \frac{(1-\alpha)^2}{2} - \alpha^2 \frac{(1-\alpha)^3}{3}\big) d\alpha} \\
&= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)}{\int_0^1 \big(\frac{1}{6}\alpha^2 - \frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^4 - \frac{1}{6}\alpha^5\big) d\alpha} \\
&= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)}{\big(\frac{1}{18}\alpha^3 - \frac{1}{8}\alpha^4 + \frac{1}{10}\alpha^5 - \frac{1}{36}\alpha^6\big)\big|_{\alpha=0}^{1}} \\
&= \frac{(\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)}{\frac{1}{360}} \\
&= 360 \cdot (\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2)
\end{aligned}
$$

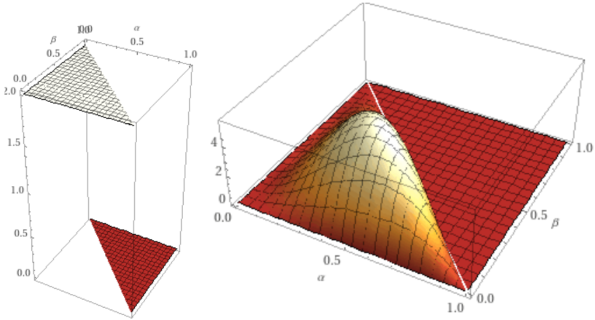
Bayes: Weather Example

Visualizing the prior and posteriors:

$$p(\alpha, \beta) = 2 \qquad \text{(prior distribution)}$$

$$p(\alpha, \beta \,|\, \mathcal{D}) = 360 \cdot (\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2) \qquad \text{(posterior distribution)}$$

Prior distribution Posterior distribution



Bayes: Weather Example

Question: What is the probability of the next events $\mathcal{D}^\star = (x_5, x_6)$ being:

rainy	rainy

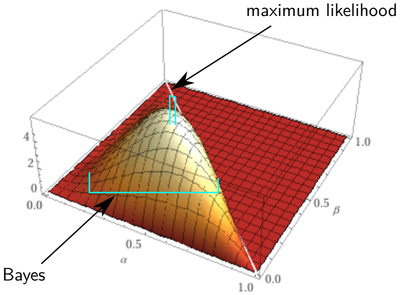
Answer:

- Recall that in the Bayesian framework, we now see the estimated parameter θ as a random variable $\Theta_\mathcal{D} \sim p(\theta \,|\, \mathcal{D})$.
- We can make the desired prediction by computing an expectation over this random variable:

$$
\begin{aligned}
\mathbb{E}[P(\mathcal{D}^\star \,|\, \Theta_\mathcal{D})] &= \int P(\mathcal{D}^\star \,|\, \theta)\, p(\theta \,|\, \mathcal{D}) d\theta \\
&= \int_0^1 d\alpha \int_0^{1-\alpha} d\beta \cdot (1-\alpha-\beta)^2 \cdot 360 \cdot (\alpha^2\beta - \alpha^3\beta - \alpha^2\beta^2) \\
&= \dots \\
&= 3/28
\end{aligned}
$$

- Recall that using maximum likelihood we obtained for the same question the different the result $1/16$.

Maximum Likelihood vs. Bayes



- Maximum likelihood only considers the most likely parameter $\theta^\star$ for making predictions.
- Bayes considers *all* parameters weighted by their probability $p(\theta \,|\, \mathcal{D})$ for making predictions.

Maximum Likelihood vs. Bayes

Maximum likelihood advantages:

- Simpler framework (no need to specify prior distributions).
- Better runtime in practice (no need for integrating probability distributions).

Bayes advantages:

- More accurate predictions are achievable, that also take into account the less likely (but still possible) parameters.

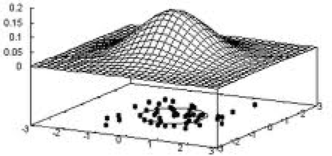
Part 3

Learning the Parameters of a Gaussian

Multivariate Gaussian Distributions

Gaussian probability density function:

$$p(\boldsymbol{x} \,|\, \theta) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$$



- Many data can be represented as vectors in $\mathbb{R}^d$.
- Gaussian distributions are a priori good general models for observations.
- Often comes with closed-form solutions.

Multivariate Gaussian Distributions

Recall that our model, assuming data to be iid. assigns to our dataset the probability:

$$p(\mathcal{D} \,|\, \theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \,|\, \theta)$$

Taking the log on both sides, we get:

$$\log p(\mathcal{D} \,|\, \theta) = \sum_{i=1}^{N} \log p(\boldsymbol{x}_i \,|\, \theta)$$

Injecting the Gaussian pdf in place of $p(\boldsymbol{x}_i \,|\, \theta)$, we get:

$$\log p(\mathcal{D} \,|\, \theta) = \sum_{i=1}^{N} -\frac{1}{2} \log \left[(2\pi)^d \det(\Sigma)\right] - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

Question:

- What are the parameters $\boldsymbol{\mu}$ and Σ that maximize the log-likelihood?

Maximum Likelihood Estimation of $\boldsymbol{\mu}$

$$\arg\max_{\boldsymbol{\mu}} \overbrace{\underbrace{\sum_{i=1}^{N} -\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})}_{J(\boldsymbol{\mu})} - \frac{1}{2}\log\left[(2\pi)^d \det(\Sigma)\right]}^{\log p(\mathcal{D}\,|\,\theta)}$$

The maximum of $J(\boldsymbol{\mu})$ is reached at a point where $\nabla J(\boldsymbol{\mu}) = \mathbf{0}$.

$$\nabla J(\boldsymbol{\mu}) = -\sum_{i=1}^{N} \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) = \mathbf{0}$$

This gives us the solution:

$$\boxed{\boldsymbol{\mu}^\star = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i}$$

Hence, among all Gaussian distributions, the one that best explains the data is the one whose mean parameter corresponds to the empirical mean of the data.

Maximum Likelihood Estimation of Σ^{-1}

Let's first make some simplifications that do not change the argmax:

$$\arg\max_{\Sigma^{-1}} \sum_{i=1}^{N} -\frac{1}{2} \log\left[(2\pi)^d \det(\Sigma)\right] - \frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= \arg\max_{\Sigma^{-1}} \underbrace{N \log\det(\Sigma^{-1}) - \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})}_{J(\Sigma^{-1})}$$

The maximum of $J(\Sigma^{-1})$ is reached at a point where $\nabla J(\Sigma^{-1}) = \mathbf{0}$.

To proceed further, we will make use of two useful identities (cf. matrix cookbook):

$$\nabla \log|\det(A)| = (A^{-1})^\top$$
$$\nabla (b^\top A b) = b b^\top$$

Maximum Likelihood Estimation of Σ^{-1} (cont.)

Recall from the previous slide that:

$$J(\Sigma^{-1}) = N \log\det(\Sigma^{-1}) - \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

Useful identities:

$$\nabla \log|\det(A)| = (A^{-1})^\top$$

$$\nabla (b^\top A b) = b b^\top$$

Taking the derivative:

$$\nabla J(\Sigma^{-1}) = N\Sigma - \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$$

and setting $\nabla J(\Sigma^{-1}) = 0$, we get the optimal parameter $\Sigma^\star$:

$$\Sigma^\star = \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$$

Injecting our maximum likelihood estimate $\boldsymbol{\mu}^\star = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$, we find that $\Sigma^\star$ is the usual empirical covariance of the data.

Maximum Likelihood Estimation of a Gaussian

Summary:

- Optimal parameters of a Gaussian distribution (that best explain the data) can be obtained in closed form.
- These optimal parameters correspond to the usual mean and covariance estimators, i.e. $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with

$$\boldsymbol{\mu}^\star = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \qquad\qquad \Sigma^\star = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}^\star)(\boldsymbol{x}_i - \boldsymbol{\mu}^\star)^\top$$

What did we gain compared to just estimating means and covariances?

- By modeling our data as a Gaussian distribution (or any distribution), we have *fully specified* the way our data is generated, and we can potentially run more complex inferences than PCA/regression/etc.

What are the risks?

- These more complex inferences are only expected to be accurate if the data is indeed Gaussian.

Part 4 **Inferences with Gaussian Distributions**

Probabilistic Model of Regression

- Assume we have $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, and we would like to predict y from $\boldsymbol{x}$ (i.e. regression).
- In our probabilistic setting, we first start by building the Gaussian density model:

$$p(\boldsymbol{x}, y) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$$

- A model of the output y given some input $\boldsymbol{x}$ can be directly obtained by the measuring conditional $p(y \,|\, \boldsymbol{x})$ of our probability model. Using the formulas for conditioning a Gaussian distribution (cf. Section 8 of the matrix cookbook), we find that this conditional distribution has the form:

$$p(y \,|\, \boldsymbol{x}) = \mathcal{N}(\mu', \Sigma')$$

with

$$\begin{aligned} \mu' &= \mu_y + (\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \Sigma_{xx}^{-1} \Sigma_{xy} \\ \Sigma' &= \Sigma_{yy} - \Sigma_{yx}^\top \Sigma_{xx}^{-1} \Sigma_{xy} \end{aligned}$$

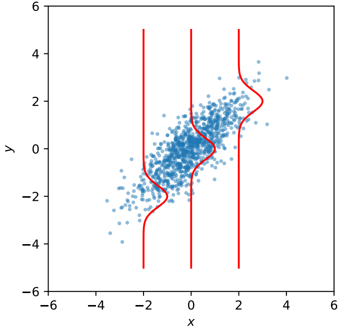
Probabilistic Models for Regression

Prediction model:

$$p(y \,|\, \boldsymbol{x}) = \mathcal{N}(\mu', \Sigma') \qquad \text{with:} \quad \begin{aligned} \mu' &= \mu_y + (\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \Sigma_{xx}^{-1} \Sigma_{xy} \\ \Sigma' &= \Sigma_{yy} - \Sigma_{yx}^\top \Sigma_{xx}^{-1} \Sigma_{xy} \end{aligned}$$

Observations:

- For each data point, we not only have a prediction, but a full distribution representing the expected value y can take. We can use this to model the error of our model.
- Notice some patterns reminiscent of least square regression, in particular, the weight $\Sigma_{xx}^{-1}\Sigma_{xy}$ of the model, and its mean square error $(\Sigma_{yy} - \Sigma_{yx}^\top \Sigma_{xx}^{-1} \Sigma_{xy})$.



Probabilistic Models for Discriminants

- In the previous lectures, we have seen different types of *linear* discriminants (e.g. difference-of-means, Fisher discriminant, support vector machines), all of them of the form $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$.
- Instead, let us now take a probabilistic approach and assume that we have as a first step built a probability model for each class:

  $$p(\boldsymbol{x} \,|\, \omega_1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$$
  $$p(\boldsymbol{x} \,|\, \omega_2) \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$$

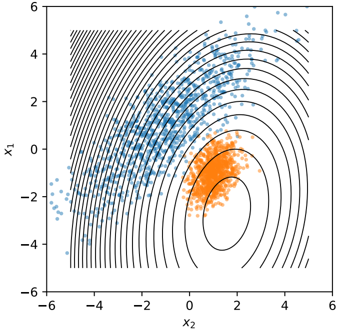
  We can now formulate the discriminant as a log-probability ratio:

  $$f(\boldsymbol{x}) = \log \frac{\overbrace{p(\boldsymbol{x} \,|\, \omega_1) \cdot P(\omega_1)/p(\boldsymbol{x})}^{P(\omega_1 \,|\, \boldsymbol{x})}}{\underbrace{p(\boldsymbol{x} \,|\, \omega_2) \cdot P(\omega_2)/p(\boldsymbol{x})}_{P(\omega_2 \,|\, \boldsymbol{x})}} = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \Sigma_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \Sigma_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) + \text{cst.}$$

  and observe that the latter is quadratic with $\boldsymbol{x}$.
- **Note:** This is an *optimal* discriminant if the probability model is correct.

Quadratic Discriminants

Example:



- Discriminant favors 'blue' any direction outside the data, because the blue distribution has generally more variance.
- This can be useful for anomaly detection, where the distribution of anomalies has typically more variations than the 'normal' data.

Quadratic Discriminants (More Examples)

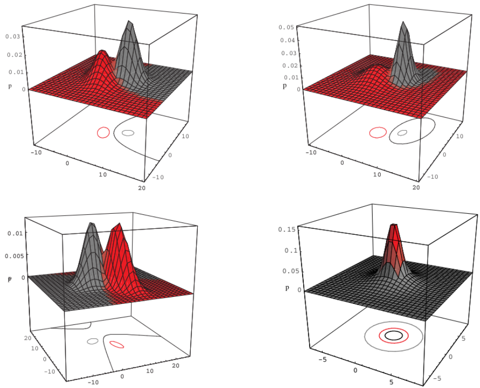


image source: Duda et al. Pattern Classification

- Discriminants can take various forms in practice, depending on the covariance structure of the two distributions (e.g. ellipses, hyperboles, etc.).

Special Cases

Recall that:

$$f(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \text{cst.}$$

if $\Sigma_1 = \Sigma_2 \overset{\text{(def)}}{=} \Sigma$ (i.e. same Gaussian distributions except for a shift), the equation reduces to the *Fisher discriminant*:

$$\begin{aligned} f(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \text{cst.} \\ &= \boldsymbol{x}^\top \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \text{cst.} \end{aligned}$$

if $\Sigma = \sigma^2 I$ (i.e. Gaussian distributions are isotropic), it further reduces to the *difference of means*:

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\sigma^2 + \text{cst.}$$

Fisher discriminants and difference of means are expected to work optimally under some *restrictive* assumptions about the class distributions. They may still be the best methods when it is not possible to get good models of Σ_1 or Σ_2, e.g. due to high dimensions and lack of data.

Summary

Summary

- Probabilistic modeling decomposes the process of building the predictive model in two steps: (1) estimating the parameters of the data-generating distribution; (2) extracting some quantity of interest from the learned probability model (e.g. a conditional mean, a likelihood ratio).
- There are two main approaches to probabilistic modeling: *Maximum likelihood* and *Bayes*. Both approaches have their advantages and limitations.
- When we use Gaussian distributions for the probability model, we may recover existing algorithms (e.g. least square regression, Fisher discriminant), but we may also get something more powerful as a result (e.g. quadratic discriminants, predictive variance).