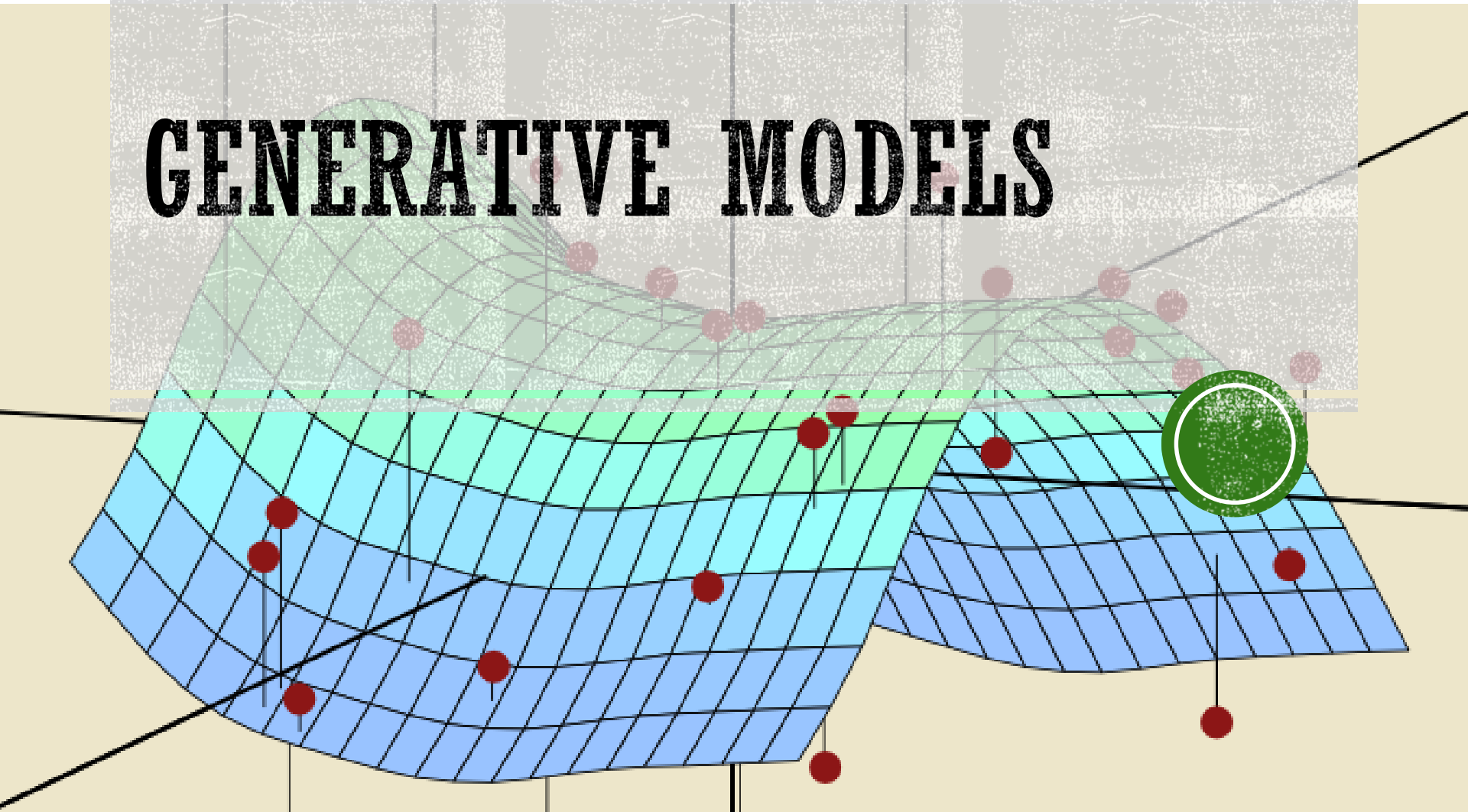
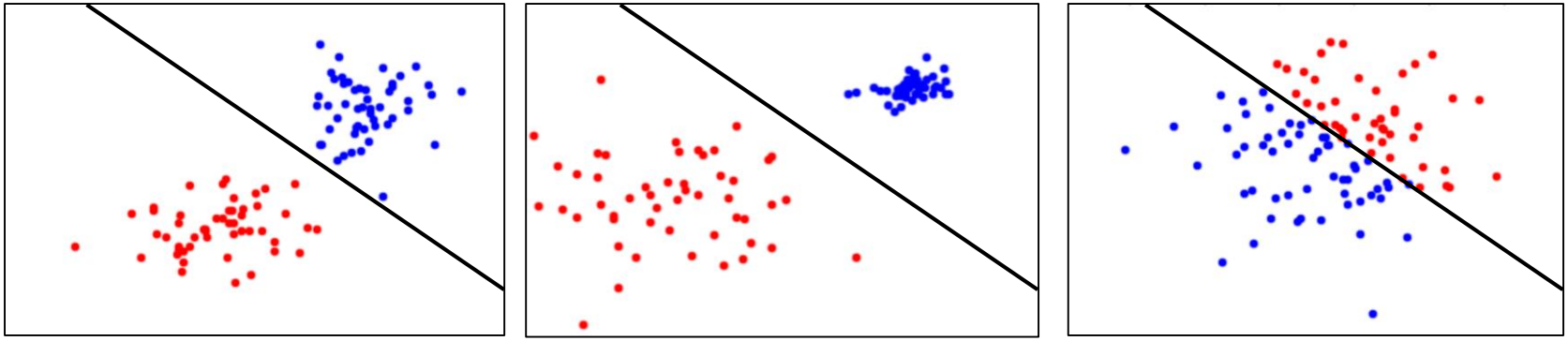
GENERATIVE MODELS

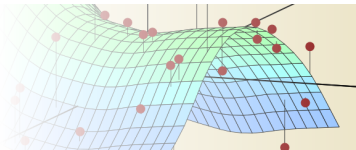


PROBABILISTIC MODELS

- Describe structure in data
- Describe process that generated the data
- Describe what we can never know, using randomness
- Understand how randomness affects our decisions



Same decision boundary. Different data statistics.

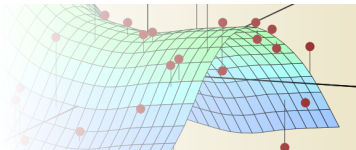


MULTINOMIAL

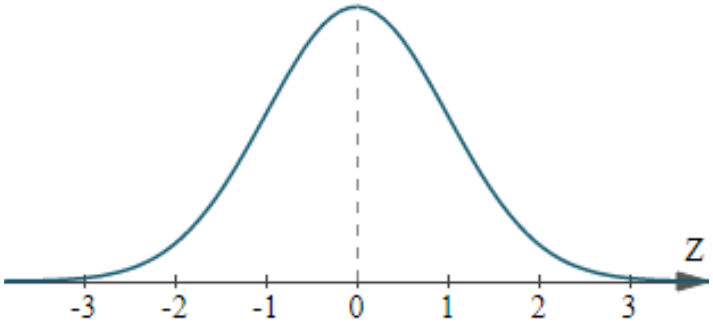
States. $x \in \{1, 2, \ldots, d\}$

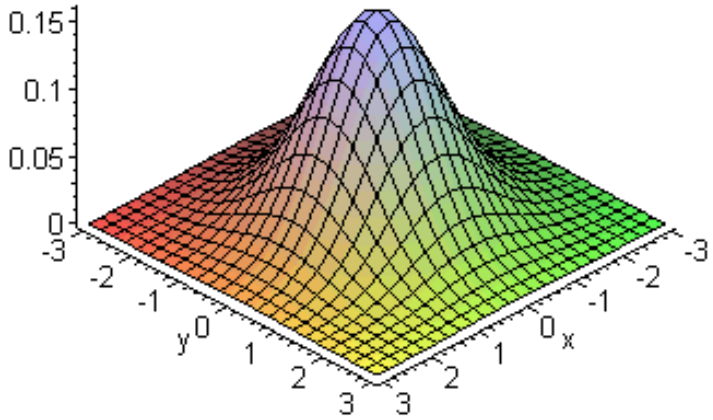
Parameters. $\theta_1, \theta_2, \ldots, \theta_d, \quad \theta_i \geq 0$ for all i, $\sum_i \theta_i = 1$.

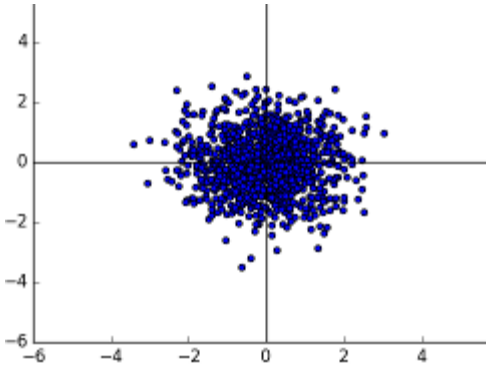
Probability Mass Function. $p(x|\theta) = \theta_x$



MULTIVARIATE GAUSSIAN

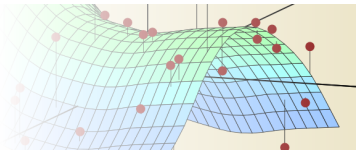






one-dimensional

two-dimensional



MULTIVARIATE GAUSSIAN

States. $x \in \mathbb{R}^d$

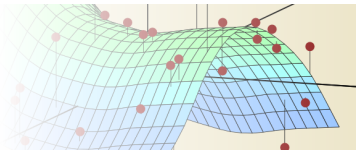
Parameters. $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$, Σ positive definite.

Probability Density Function.

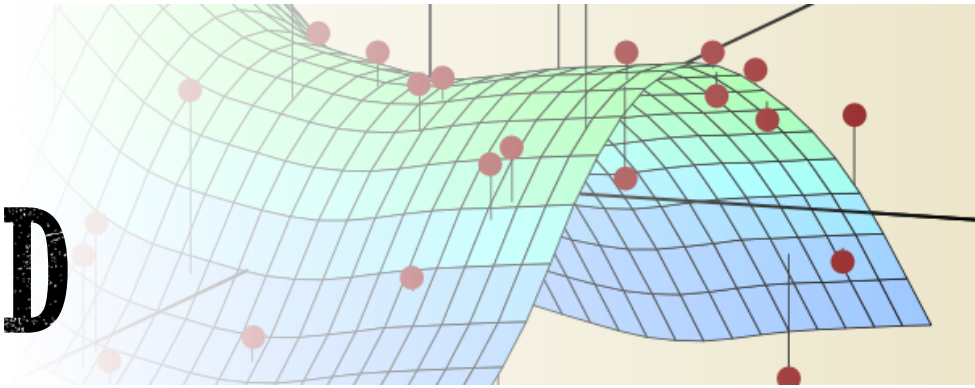
$$p(x \mid \mu, \Sigma) = (2\pi)^{-d/2} (\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

Spherical Gaussian.

$$p\left(x \mid \mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2}\|x-\mu\|^2\right\}, \quad \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}.$$



MAXIMUM LIKELIHOOD



MAXIMUM LIKELIHOOD ESTIMATE (MLE)

PMF or PDF

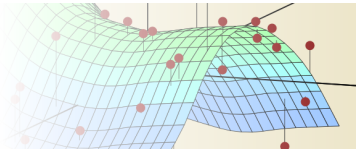
Model.
For each parameter $\theta \in \mathbb{R}^d$, we have a distribution $p(x|\theta)$.

Data.
Set of observations $\mathcal{S} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$

Parameter estimation.
Find the parameter $\hat{\theta} \in \mathbb{R}^d$ that 'best' describes the data $\mathcal{S}$.

Likelihood. $p(\mathcal{S}|\theta) = \prod_{x \in \mathcal{S}} p(x|\theta)$

Maximum Likelihood Estimate. $\hat{\theta} = \text{argmax}_\theta \, p(\mathcal{S}|\theta)$



LOG LIKELIHOOD

Maximizing the likelihood

$$p(\mathcal{S}|\theta) = \prod_{x\in\mathcal{S}} p(x|\theta)$$

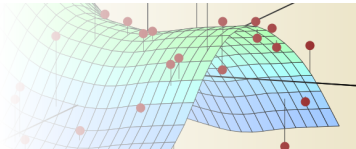
is the same as minimizing the negative log likelihood

$$-\frac{1}{\mathrm{n}}\log p(\mathcal{S}|\theta) = -\frac{1}{n}\log\prod_{x\in\mathcal{S}} p(x|\theta)$$

$$= \frac{1}{n}\sum_{x\in\mathcal{S}} -\log p(x|\theta).$$

In fact, we will define this to be the *training loss.*

It is the average of the point loss $-\log p(x|\theta)$.



MLE – MULTINOMIAL

Example.

Treat a document $\mathcal{S}$ as a *bag of words*, counting only how many times $n(w)$ each dictionary word $w \in W$ appears in $\mathcal{S}$.

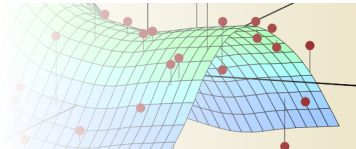
Assume that the document was *generated* one word at a time, each word independently from the others.

Assume that words are drawn from same distribution $\theta = (\theta_w)$ where $w \in W$ appears with probability θ_w.

independent and identically distributed

Likelihood. $p(\mathcal{S}|\theta) = \prod_{w \in W} \theta_w^{n(w)}$

Training Loss. $\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{w \in W} -n(w) \log \theta_w$



MLE – MULTINOMIAL

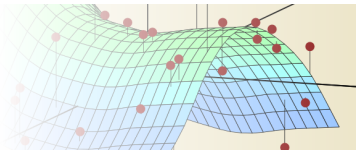
minimize $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{w \in W} -n(w) \log \theta_w$

subject to $\theta_w \geq 0$ for all $w \in W$

$\sum_{w \in W} \theta_w = 1$

Using Lagrange multipliers, we showed that the minimum is attained at the following point.

MLE. $\hat{\theta}_w = \frac{n(w)}{\sum_{w' \in W} n(w')}$



MLE – GAUSSIAN

Example.

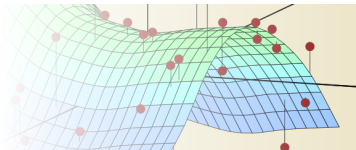
Assume that data $\mathcal{S} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ is independent and identically distributed to a spherical Gaussian with mean $\mu \in \mathbb{R}^d$ and variance σ^2.

Training Loss.

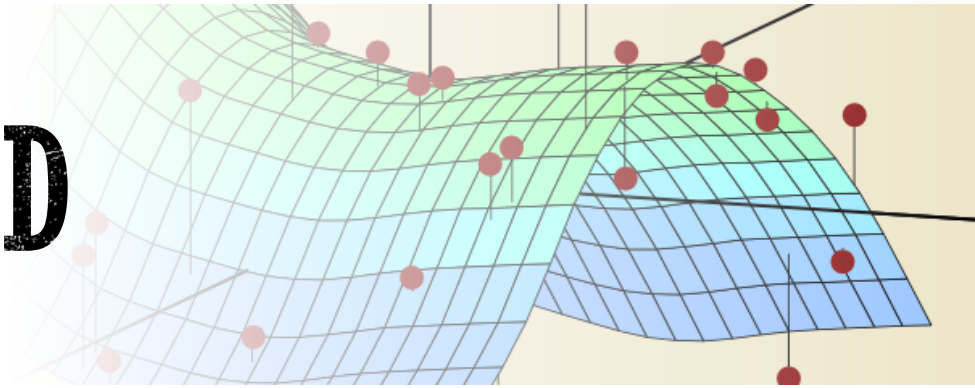
$$\mathcal{L}_n(\mu, \sigma^2) = -\frac{1}{n}\log p(\mathcal{S}|\mu, \sigma^2) = -\frac{1}{n}\sum_{x\in\mathcal{S}} \log p(x|\mu, \sigma^2)$$
$$= \frac{d}{2}\log(2\pi\sigma^2) + \frac{1}{2n\sigma^2}\sum_{x\in\mathcal{S}}\|x-\mu\|^2$$

MLE. $\hat{\mu} = \frac{1}{n}\sum_{x\in\mathcal{S}} x\,, \quad \hat{\sigma}^2 = \frac{1}{nd}\sum_{x\in\mathcal{S}}\|x-\hat{\mu}\|^2$

solve
$\frac{\partial \ell}{\partial \mu} = 0$
$\frac{\partial \ell}{\partial \sigma} = 0$



LOG LIKELIHOOD RATIO



CLASSIFICATION - MULTINOMIAL

Example.

Doc labels $+, -$. Bags of words $\mathcal{S}^+, \mathcal{S}^-$. New unlabeled doc $\mathcal{S}$.

First, compute MLEs for the likelihoods

$$p(\mathcal{S}^+|\theta^+) = \prod_{w \in \mathcal{S}^+} \theta_w^+, \qquad p(\mathcal{S}^-|\theta^-) = \prod_{w \in \mathcal{S}^-} \theta_w^-.$$

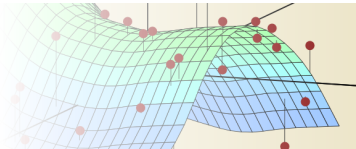
Likelihood ratio.

We classify $\mathcal{S}$ by checking the sign of the log likelihood ratio

$$\log \frac{p(\mathcal{S}|\theta^+)}{p(\mathcal{S}|\theta^-)} = \sum_w n(w) \log \frac{\theta_w^+}{\theta_w^-}.$$

θ_w

Linear classifier!



CLASSIFICATION - MULTINOMIAL

Bayesian posterior.

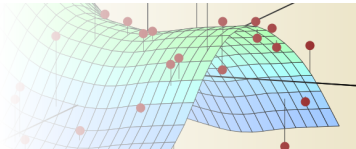
Assume *prior probabilities* $p(+), p(-)$ on document labels.
Again, we classify using the sign of the log likelihood ratio

$$\log \frac{p(\mathcal{S}|\theta^+)p(+)}{p(\mathcal{S}|\theta^-)p(-)} = \sum_w n(w) \underbrace{\log \frac{\theta_w^+}{\theta_w^-}}_{\theta_w} + \underbrace{\log \frac{p(+)}{p(-)}}_{\theta_0}.$$

Linear classifier with offset!

This is equivalent to choosing the label with the higher *posterior probability*

$$p(+|\mathcal{S}) = \frac{p(\mathcal{S}|+)p(+)}{p(\mathcal{S})} = \frac{p(\mathcal{S}|\theta^+)p(+)}{p(\mathcal{S}|\theta^+)p(+)+p(\mathcal{S}|\theta^-)p(-)}$$



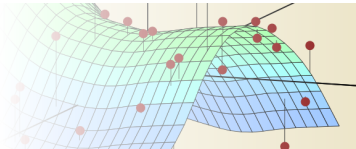
CLASSIFICATION – GAUSSIAN

Example.

Classifying between two Gaussians with the same variance σ^2 but different means μ^+, μ^-.

$$\begin{aligned}
\log \frac{P(x|\mu^+,\sigma)}{P(x|\mu^-,\sigma)} &= \log P(x|\mu^+,\sigma) - \log P(x|\mu^-,\sigma) \\
&= \log[C \cdot e^{-\frac{1}{2\sigma^2}||x-\mu^+||^2}] - \log[C \cdot e^{-\frac{1}{2\sigma^2}||x-\mu^-||^2}] \\
&= \frac{1}{2\sigma^2}(2x \cdot \mu^+ - 2x \cdot \mu^-) - ||\mu^+||^2 + ||\mu^-||^2 \\
&= \underbrace{\frac{x \cdot (\mu^+ - \mu^-)}{\sigma^2}}_{\theta} - \underbrace{\frac{1}{2\sigma^2}(||\mu^+||^2 - ||\mu^-||^2)}_{\theta_0}
\end{aligned}$$

Linear classifier with offset!

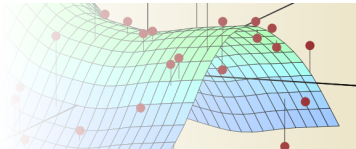


SMOOTHING – MULTINOMIAL

Problem.

What if training data does not include word w? Then, $\hat{\theta}_w = 0$.
What if test data includes word w? Then, classification fails.

$$\log \frac{p(\mathcal{S}|\theta^+)}{p(\mathcal{S}|\theta^-)} = \sum_w n(w) \log \frac{\theta_w^+}{\theta_w^-}$$



SMOOTHING – MULTINOMIAL

e.g. choose λ to be small but proportional to word distribution in all books

Solution.

Introduce prior distribution over θ.

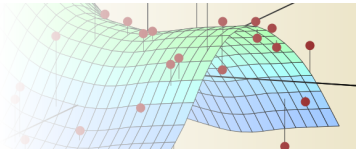
$$P(\theta|\lambda_1, \dots, \lambda_n) = P(\theta|\lambda) = C \prod_{w \in W} \theta_w^{\lambda_w}$$

Estimate parameters by maximizing the posterior probability.

$$P(\theta|S) \propto P(S|\theta)P(\theta|\lambda) = \prod_{w \in W} \left[\theta_w^{n(w)} \cdot \theta_w^{\lambda_w}\right] = \prod_{w \in W} \left[\theta_w^{n(w)+\lambda_w}\right]$$

The hyper-parameters $\lambda_1, \dots, \lambda_n > 0$ act as *pseudo-counts*.

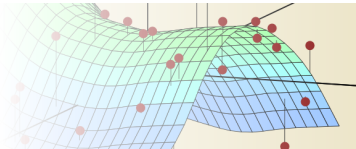
They give default values to θ_w if w does not appear in training, and their effect diminishes if w appears often in training.



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

SUMMARY

- Distributions
  - Multinomial
  - Multivariate Gaussian
  - Spherical Gaussian
- Maximum Likelihood
  - Training Loss
  - Multinomial MLE
  - Gaussian MLE
- Log Likelihood Ratio
  - Multinomial classifier
  - Gaussian classifier
- Smoothing
  - Prior



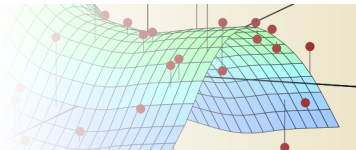
INTENDED LEARNING OUTCOMES

Probabilistic Models

- Explain how probabilistic modeling is useful in machine learning.
- Define the multinomial and the multivariate Gaussian distributions.
- Define *independent and identically distributed* (i.i.d.)

Maximum Likelihood

- Given a probabilistic model and data, describe a training loss for parameter estimation. Define maximum likelihood estimate (MLE).
- Given a probabilistic model and data, derive the MLE. Write down the MLE for multinomial and multivariate Gaussian distributions.



INTENDED LEARNING OUTCOMES

Log Likelihood Ratio

- Describe how to use the log likelihood ratio for classification.
- Derive linear classifiers given that the distribution for each label class is multinomial or multivariate Gaussian.

Bayesian Methods

- Describe how to compute the posterior probabilities using the prior probabilities and model probabilities.
- Explain how priors $p(+), p(-)$ on labels give classifiers with offset.
- Explain how priors $p(\theta)$ on parameters give smoothed estimates.