Introduction to Machine Learning.
CSCI-UA 9473, Lecture 2.

Augustin Cosse

Ecole Normale Supérieure, DMA & NYU
Fondation Sciences Mathématiques de Paris.

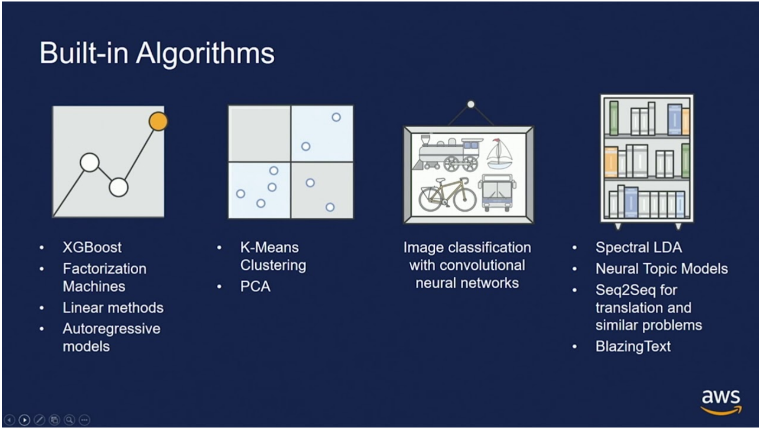
2018

Previous lecture

- General overview
  - What is Machine Learning?
  - How does it fit between deep Learning and Artificial Intelligence?
  - What are the different classes of methods?
- What are the risks ?
- What are the challenges?
- What are we going to do during the class ?

Remember the algorithms we are going to study?

From CWS 2018, decision making with Amazon SageMaker



Built-in Algorithms

- XGBoost
- Factorization Machines
- Linear methods
- Autoregressive models

- K-Means Clustering
- PCA

Image classification with convolutional neural networks

- Spectral LDA
- Neural Topic Models
- Seq2Seq for translation and similar problems
- BlazingText

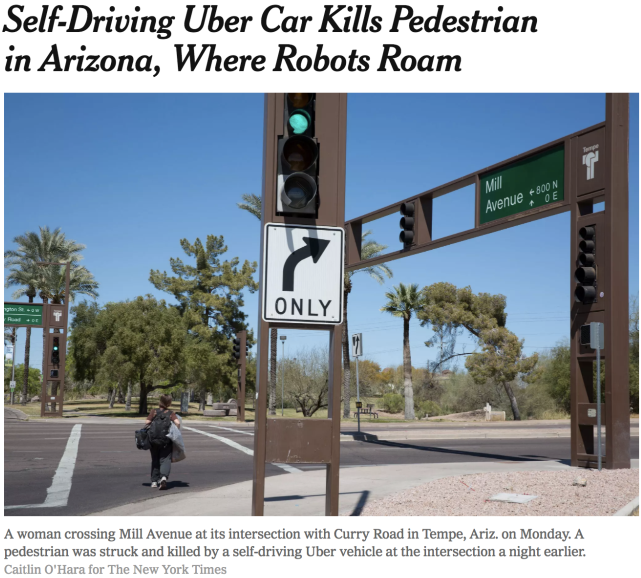
Today

- General reminders on statistics and probability
- Bayesian vs Frequentist
- First (short) intro to online (programming) tools

Statistics and probability

- Why using stats/proba?
- Machine Learning relies on complex distributions (cancerous cells, possible moves in Go, Existing sign roads, possible evolutions of stocks, connections between people, words,..)
- Only a few samples are usually available
- $\Rightarrow$ We need a way to measure how well those samples are representing the underlying (unknown) distribution

Why is that important?



Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam

A woman crossing Mill Avenue at its intersection with Curry Road in Tempe, Ariz. on Monday. A pedestrian was struck and killed by a self-driving Uber vehicle at the intersection a night earlier.

Caitlin O'Hara for The New York Times

Alex Kendall

Computer Vision &
Robotics Researcher

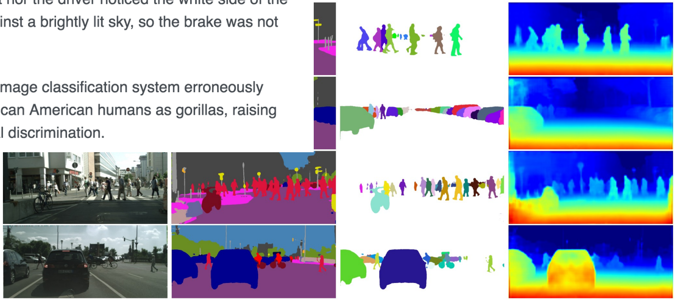
University of Cambridge

Deep Learning Is Not Good Enough, We Need Bayesian Deep Learning for Safe AI

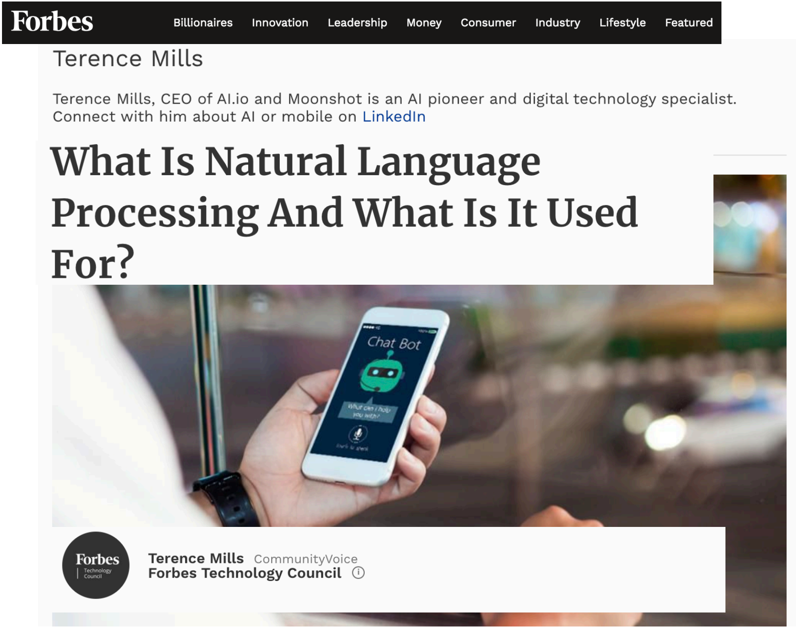
Bayesian Deep Learning, Computer Vision, Uncertainty

1. In May 2016 we tragically experienced the first fatality from an assisted driving system. According to the manufacturer’s blog, “Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied.”

2. In July 2015, an image classification system erroneously identified two African American humans as gorillas, raising concerns of racial discrimination.



(a) Input image (b) Segmentation output (c) Instance output (d) Depth output



Terence Mills

Terence Mills, CEO of AI.io and Moonshot is an AI pioneer and digital technology specialist. Connect with him about AI or mobile on LinkedIn

What Is Natural Language Processing And What Is It Used For?

Terence Mills CommunityVoice
Forbes Technology Council ⓘ

Reminders (I)

(Discrete sets of events)

- Sum rule $p(X) = \sum_Y p(X|Y)$
- Product rule $p(X, Y) = p(X|Y)p(Y)$
- Bayes theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

(continuous sets of events)

- density $p(x)$, marginalizing

$$p(x \in [a, b]) = \int_a^b p(x)dx, \quad p(x) = \int p(x, y)dy$$

Reminders (II)

- Cumulative distribution Function (CDF) $F(z) = \int_{-\infty}^{z} p(x)\, dx$
- Expectation $\mathbb{E}[x] = \int xp(x)dx$, $\mathbb{E}[x] = \sum_i x_i p(x_i)$
- Conditional expectation $\mathbb{E}_x f(x|y) = \sum_x f(x)p(x|y)$
- Variance $\mathrm{Var}[x] \equiv \mathbb{E}\left\{(x - \mathbb{E}x)^2\right\}$
- Covariance $\mathrm{Cov}[x, y] \equiv \mathbb{E}\left\{(x - \mathbb{E}x)(y - \mathbb{E}y)\right\}$

Reminders (III) A few important distributions

- The gaussian distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- Uniform distribution: $P(y) = \frac{1}{|b-a|}, \quad y \in [a,b]$

- χ^2 distribution: $\chi^2 \sim \sum_{i=1}^{N} {Z_i}^2$ with Z_i independent standard normal RV.

Reminders (IV) A few important distributions

- Binary variables: Bernoulli and Rademacher,

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}, \quad x = \left\{ \begin{array}{c} 1 \\ 0 \end{array} \right., 0 \leq \mu \leq 1$$

$$\text{Rademacher:} \quad \varepsilon(x) = \begin{cases} (1/2), & x = +1 \\ (1/2), & x = -1 \\ 0, & \text{otherwise} \end{cases}$$

Reminders (IV) A few important distributions

- Beta distribution

$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{(a-1)}(1-x)^{b-1}$$

$$B(a,b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

- $\Gamma(a)$ is the Gamma function.

The exponential family

- Many of the distributions we have discussed are part of a general family called The exponential family
- The exponential family has many interesting properties
  - It is the only family of distribution with finite-sized sufficient statistics
  - It is the only family with known conjugate priors
  - It is at the core of generalized linear models
  - it is at the core of variational inference
- we will come back to these notions later

The exponential family

- A pdf $p(\boldsymbol{x}|\theta)$ is said to be in the exponential family for $\boldsymbol{x} = (x_1, \ldots, x_m)$ and $\theta \subseteq \mathbb{R}^d$ if

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp(\theta^T \phi(\boldsymbol{x}) \\
&= h(\boldsymbol{x}) \exp(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta}))
\end{aligned}
$$

- $Z(\theta)$ and $A(\theta)$ are defined as

$$
\begin{aligned}
Z(\theta) &= \int_{\mathcal{X}^m} h(\boldsymbol{x}) \exp[\theta^T \phi(x)]\, dx \\
A(\theta) &= \log(Z(\theta))
\end{aligned}
$$

- $Z(\theta)$ is called the partition function, θ are the mutual parameters, $\phi(x) \in \mathbb{R}^d$ is the vector of sufficient statistics, $A(\theta)$ is the log partition function or cumulant function.

The exponential family

- Two examples
  - Bernoulli

    $$\begin{aligned}\mathrm{Ber}(x|\mu) = \mu^x(1-\mu)^{1-x} &= \exp(x\log(\mu) + (1-x)\log(1-\mu)) \\ &= \exp(\phi(x)^T\theta)\end{aligned}$$

  - Univariate Gaussian
- The Uniform distribution does not belong to the exponential family

Parameter/model inference: Bayesian vs frequentist

- The linear regression model is a special instance of a more general idea called model inference (among which one finds the MLE)
- We will study the notion of inference in more details later in the class. For now we only cover the main ideas.
- Inference can be used in both supervised (learn new labels from training labels) and unsupervised (learn parameters from distribution) frameworks
- You will often hear about frequentist vs Bayesian approaches.

Parameter/model inference: Bayesian vs frequentist

- Bayesian statistics.
  - Considers the (distribution) parameters as random
  - Relies heavily on the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$
  - dominated statistical practice before 20^{th} century
  - Ex: MAP $\underset{\theta}{\operatorname{argmax}} P(\mathcal{D}|\theta)P(\theta)$
- Frequentist statistics (a.k.a classical stat.)
  - Parameters $\boldsymbol{\theta}$ viewed as fixed, sample $\mathcal{D}$ as random (Randomness in the data affects the posterior)
  - Relies on the likelihood or some other function of the data
  - dominated statistical practice during 20^{th} century
  - Ex. MLE : $\underset{\theta}{\operatorname{argmax}} P(\mathcal{D}|\theta)$

Bayesian statistics: Some vocabulary

- We saw Bayesian inference relies on the posterior $p(\theta|\mathcal{D})$
- The posterior reads from the Bayes rule as

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &= \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta} \end{aligned}$$

- $p(\theta)$ is called the prior, $p(\mathcal{D}|\theta)$ is called the likelihood function and $Z = p(\mathcal{D})$ is the normalizing constant (independent of θ)
- Given a set of patterns $(\boldsymbol{x}_\mu, \boldsymbol{y}_\mu)$, classifiers are usually of two types:
    - Generative (learn model for $p(\boldsymbol{x}, \boldsymbol{y}|\theta)$)
    - Discriminative (learn model for $p(\boldsymbol{y}|\boldsymbol{x}, \theta)$)

Bayesian statistics: Some vocabulary

- An example of discriminative classifier : Logistic regression
  - Here we take $\mu(\boldsymbol{x}) = \text{sigm}(\boldsymbol{w}^T\boldsymbol{x})$ and define the classifier as a Bernoulli distribution

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \text{Ber}(y|\mu(\boldsymbol{x}))$$

  - Good when the output is binary

- An example of generative classifier : Naive Bayes
  - relies on the assumption that the features (hidden variables) are independent

$$p(\boldsymbol{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} p(x_j|y = c, \theta_{jc})$$

  - $\theta_{j,c}$ is the parameters of the distribution of class c for j^{th} entry in the D-dimensional pattern vector $\boldsymbol{x} \in \{1, \ldots, K\}^D$.

- We will study those models in further detail when discussing classifiers.

Bayesian statistics

- In Bayesian statistics, randomness is most often used to encode uncertainty
- The posterior $p(\boldsymbol{\theta}|\mathcal{D})$ summarizes all we know on the parameters
- Bayesian inference is not always the right choice because of the following
  - The Mode is not a typical point in the distribution
  - MAP estimator depends on re-parametrization

Bayesian statistics: Drawbacks and solutions

- A solution to the first part is to use a more robust loss function $\ell(\hat{\theta}, \theta)$
- A solution to the second part is to replace the MAP with the following estimator (when available)

$$\hat{\theta} = \underset{\theta}{\mathrm{argmax}} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) |\boldsymbol{I}(\boldsymbol{\theta})|^{-1/2} \tag{1}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is the Fischer information matrix

Fischer information matrix

- For a generative model $p(\boldsymbol{x}|\theta)$, we let $g(\theta, \boldsymbol{x})$ denote the Fischer score

$$g(\theta, \boldsymbol{x}) = \nabla_\theta \log(p(\boldsymbol{x}|\theta))$$

- the Fischer Kernel is the defined as

$$k(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{\theta}, \boldsymbol{x})^T \boldsymbol{F}^{-1} g(\boldsymbol{\theta}, \boldsymbol{x}')$$

- The matrix $\boldsymbol{F}$ is called the Fischer matrix and defined as

$$\boldsymbol{F} = \mathbb{E}_{\boldsymbol{x}} \left\{ g(\boldsymbol{\theta}, \boldsymbol{x}) g(\boldsymbol{\theta}, \boldsymbol{x})^T \right\}$$

- Note that it is often computed empirically as

$$\boldsymbol{F} \approx \frac{1}{N} \sum_{n=1}^{N} g(\theta, \boldsymbol{x}) g(\boldsymbol{\theta}, \boldsymbol{x})^T$$

Occam's razor and Model selection

- Only looking for the best model often leads to overfitting (we will see that later in more details)

- Bayesian framework offers and alternative called Bayesian model selection

- For a family of models, we can express the posterior

$$
\begin{aligned}
p(m|\mathcal{D}) &= \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m|\mathcal{D})} \\
&\propto p(\mathcal{D}|m)p(m)
\end{aligned}
$$

where $p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$ is called the marginal likelihood, integrated likelihood or evidence

Occam's razor

- Integrating the parameters $\boldsymbol{\theta}$ such as in

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

  acts as a natural regularization and prevents overfitting when solving for $\max_m p(m|\mathcal{D})$. This idea is known as Bayesian Occam's razor

- The evidence $p(\mathcal{D}|m)$ can be understood as the probability to generate a particular dataset from a family of model (all values of the parameters included).

- When the family of models is too simple, or too complex, this probability will be low.

Bayesian decision theory

- How do we resolve the lack of robustness of Bayesian estimators vis a vis the distribution (recall the bimodal distribution)?
- Statistical decision theory can be viewed as a game against nature.
- Nature has a parameter value in mind and gives us a sample
- We then have to guess what the value of the parameter is by choosing an action a
- As an additional piece of information, we also get a feedback from a loss function $L(y, a)$ which tells us how compatible our action is w.r.t Nature's hidden state.

Bayesian decision theory

- The goal of the game is to determine the optimal decision procedure,

$$\underset{a\in\mathcal{A}}{\operatorname{argmin}}\,\mathbb{E}\left\{L(y,a)\right\}$$

- In economics $L(y,a) = U(y,a)$ and leads to the Maximum utility principle which is considered as rational behavior

$$\delta(x) = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\,\mathbb{E}\left\{U(y,a)\right\}$$

- In the Bayesian framework, we want to minimize the loss over the models compatible with the observations $\{\boldsymbol{x}_\mu\}$

$$\delta(x) = \underset{a\in\mathcal{A}}{\operatorname{argmin}}\mathbb{E}_{p(\theta|\{\boldsymbol{x}_\mu\})}\left\{L(\theta,a)\right\} = \sum_{\theta\in\Theta} L(\theta,a)p(\theta|\left\{\boldsymbol{x}_\mu\right\}_\mu)$$

Bayesian decision theory (continued)

- The MAP is equivalent to minimizing a 0/1 loss

$$L(\hat{\theta},\theta) = \mathbb{1}_{\theta\neq\hat{\theta}} = \begin{cases} 0 & \text{if } \hat{\theta}\neq\theta \\ 1 & \text{if } \hat{\theta}=\theta. \end{cases}$$

we then have

$$\begin{aligned} \mathbb{E}L(\hat{\theta},\theta) = p(\theta\neq\hat{\theta}|\left\{\boldsymbol{x}_\mu\right\}_\mu) &= 1 - p(\hat{\theta}=\theta|\left\{\boldsymbol{x}_\mu\right\}_\mu) \\ &= 1 - p(\hat{\theta}=\theta|\left\{\boldsymbol{x}_\mu\right\}_\mu)p(\theta|\boldsymbol{x}_\mu) \end{aligned}$$

which is maximized for $\hat{\theta}=\theta$ with θ taken as

$$\theta^*(\left\{\boldsymbol{x}_\mu\right\}_\mu) = \underset{\hat{\theta}}{\operatorname{argmax}}\; p(\theta|\left\{\boldsymbol{x}_\mu\right\}_\mu)$$

What do we do with noisy data?

- Is it possible to take more robust losses?
- ℓ_2 loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|^2$ gives posterior mean

$$\mathbb{E}\left\{(\hat{\theta} - \theta)^2 | \boldsymbol{x}_\mu\right\} = \mathbb{E}[\theta^2 | \boldsymbol{x}_\mu] - 2\hat{\theta}\mathbb{E}[\theta | \boldsymbol{x}_\mu] + \hat{\theta}^2$$

- setting derivative to 0, $\partial_{\hat{\theta}} \mathbb{E}\{\hat{\theta} | \boldsymbol{x}_\mu\} = 0$, we get

$$-2\mathbb{E}\left\{\theta | \boldsymbol{x}_\mu\right\} + 2\hat{\theta} = 0$$

$$\hat{\theta} = \int \theta p(\theta | \boldsymbol{x}_\mu) \, d\theta$$

What do we do with noisy data? (continued)

- Is it possible to take more robust losses?
- ℓ_1 loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ gives posterior median
- The value $\hat{\theta}$ such that

$$p(\theta < \hat{\theta} | \boldsymbol{x}_\mu) = p(\theta \geq \hat{\theta} | \boldsymbol{x}_\mu) = 1/2$$

What do we do with noisy data? (continued)

- Now assume $\hat{\theta}$ defines the value of some hidden variable y (e.g. the class of a point $\boldsymbol{x}_\mu$ defined by a gaussian mixture $\hat{\boldsymbol{\theta}}$).
- Finding the optimal parameters (or equivalently estimate the hidden state) can be done by considering the error

$$\begin{aligned} L_g(\theta, \hat{\theta}) &= \mathbb{E}_{(\boldsymbol{x}_\mu, y_\mu)\sim p(\boldsymbol{x}_\mu, y_\mu|\theta)} \left\{ \ell(\theta, f_{\hat{\theta}}) \right\} \\ &= \sum_{\boldsymbol{x}_\mu} \sum_{y_\mu} \ell(y_\mu, f_{\hat{\theta}}(x_\mu)) p(x_\mu, y_\mu | \boldsymbol{\theta}) \end{aligned}$$

- The Bayesian approach then minimizes the posterior expected loss

$$\operatorname*{argmin}_{\hat{\theta}} \int p(\theta|\mathcal{D}) L_g(\theta, \hat{\theta})\, d\theta$$

- Note that here the model is fixed and we want to learn the parameters ($><$ model selection)

How to pick up the priors?

- The controversial aspect of Bayesian statistics are the priors
- The main argument of Bayesians is that we most often know something about the world
- When it is possible, it makes things easier to pick up a prior from the same family as the likelihood function
- Another choice is to use uninformative priors

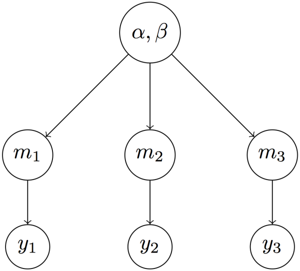
Hierarchical Bayes (I)

- Sometimes we want to use several levels of (hyper-)parameters

$$\mathcal{D} \leftarrow \boldsymbol{m} \leftarrow (\alpha, \beta)$$

- The resulting model is known as Hierarchical Bayes or multi-level model

- E.g.: related cancer rates



Hierarchical Bayes (II)

- We want to predict cancer rates in various cities. Suppose we measure the number of people in each city i, N_i, and the number of cancers x_i
- we assume that the number of cancers follows a Binomial, $x_i = \text{Bin}(N_i, \theta_i)$
- We could assume that all rates are independent or all the same $\theta_i = \theta$ for all i. An alternative is to assume θ_i are similar but with some inter city variation

$$p(\mathcal{D}, \theta, \eta | N) = p(\eta) \prod_{i=1}^{N} \text{Bin}(x_i | N_i, \theta_i) \text{Beta}(\theta_i, \boldsymbol{\eta})$$

Empirical Bayes

- In graphical models, we will need to compute the posterior on multiple levels of latent variables
- Imagine a two level posterior $p(\eta, \theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta|\eta)p(\eta)$
- Sometimes it is possible to get rid of the first order parameters by marginalizing (whenever we can compute the integral)
- then we can estimate the second order parameters as

$$\hat{\boldsymbol{\eta}} = \underset{\eta}{\mathrm{argmax}} p(\mathcal{D}|\eta) = \underset{\eta}{\mathrm{argmax}} \left[\int p(\mathcal{D}|\theta)p(\theta|\eta)d\theta \right]$$

- This approach is called Empirical Bayes or Type II maximum likelihood (in ML sometimes called the evidence procedure)

The Frequentist approach

- Unlike the Bayesian framework, the frequentist framework does not treat the parameters as random variables

- Frequentist statistics thus does not rely on the use of priors

- In frequentist statistics, randomness occurs from the variation across multiple trials

- The estimator is computed by applying some function to the data,

$$\hat{\theta} = \delta(\mathcal{D})$$

The Frequentist approach

- Uncertainty is encoded in the sampling distribution.
- Assume that we have access to S data sets generated from $p(\cdot, \theta^*)$
- for each dataset, we get one estimator $\hat{\theta}(\mathcal{D}^s)$, $s = 1, \ldots, S$
- When taking $S \to \infty$, the distribution we obtain is called sampling distribution.
- The sampling distribution is the distribution encoding the uncertainty on the estimator

Frequentist decision theory

- In Frequentist statistics there is a loss and a probability distribution
- but there is no prior and hence, no posterior
- To select an estimator, the frequentist framework considers the expected loss

$$R(\theta^*, \delta) = \mathbb{E}_{p(\tilde{\mathcal{D}}|\theta^*)}\left[L(\theta^*, \delta(\tilde{\mathcal{D}}))\right] = \int L(\theta^*, \delta(\tilde{\mathcal{D}}))p(\tilde{\mathcal{D}}|\theta^*)d\tilde{\mathcal{D}}$$

- The problem is that we do not have access to θ^*..

Frequentist Decision theory: possible fixes

- One way would be to nevertheless consider a posterior on θ^*.
- In this case we could write

$$R(\delta) = \mathbb{E}_{p(\theta^*)}[R(\theta^*, \delta)] = \int R(\theta^*, \delta) p(\theta^*) d\theta^*$$

- One estimator (which then becomes Bayesian) is then given by

$$\delta_B = \operatorname*{argmin}_{\delta} R_B(\delta)$$

- The problem is that in the Frequentist setting , we don't want to do Bayesian stats

Frequentist Decision theory: possible fixes

- An alternative is to define the maximum risk

$$R_{\max}(\delta) \equiv \max_{\theta^*} R(\theta^*, \delta)$$

- The minimax estimator is then defined as

$$\delta_{MM} \equiv \operatorname*{argmin}_{\delta} R_{\max}(\delta)$$

How to choose the estimators optimally?

- Frequentist decision theory does not provide a way to choose the best estimator
- There are however desirable properties that we will usually want the estimator to have
  - Consistency. An estimator is consistent if

$$\hat{\theta}(\mathcal{D}) \to \theta^* \quad \text{when } |\mathcal{D}| \to \infty$$

  - An estimator is said to be unbiased if

$$\text{bias}(\hat{\theta}(\cdot)) = \mathbb{E}_{p(\mathcal{D}|\theta^*)} \left\{ \hat{\theta}(\mathcal{D}) - \theta^* \right\} = 0$$

  - An estimator is assymptotically optimal if it achieves the smallest assymptotic variance among all unbiased estimators

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta^*)} \quad \text{Cramer-Rao bound}$$

How to choose the estimators optimally?

- On more thing..
- Although we cannot really choose an optimal estimator within the frequentist framework because we don't have access to the true parameter $\hat{\theta}$
- However there are estimators which are always better than other whatever be the value of θ^*
- When two estimators δ_1 and δ_2 are such that

$$R(\theta^*, \delta_1) \leq R(\theta^*, \delta_2), \quad \text{for all } \theta^*$$

- We will say that δ_1 dominates δ_2
- Finally we say that $\hat{\theta}$ is admissible if it is not strictly dominated by any other estimator

How about the MLE?

- The MLE is a consistent estimator (corresponds to minimizing $\mathbb{KL}(p(\cdot, \theta^*), p(\cdot, \hat{\theta})))$
- How about the bias ? Let's check the MLE for the Gaussian distribution

$$\text{Mean} \ : \ \text{bias}(\hat{\mu}) = \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N} x_i - \mu\right\} = 0$$

$$\text{Variance} \ : \ \mathbb{E}\hat{\sigma}^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2 = \frac{N-1}{N}\sigma^2$$

- Although the MLE may be biased, it is assymptotically unbiased
- MLE meets the Cramer-Rao bound and is this assymptotically unbiased

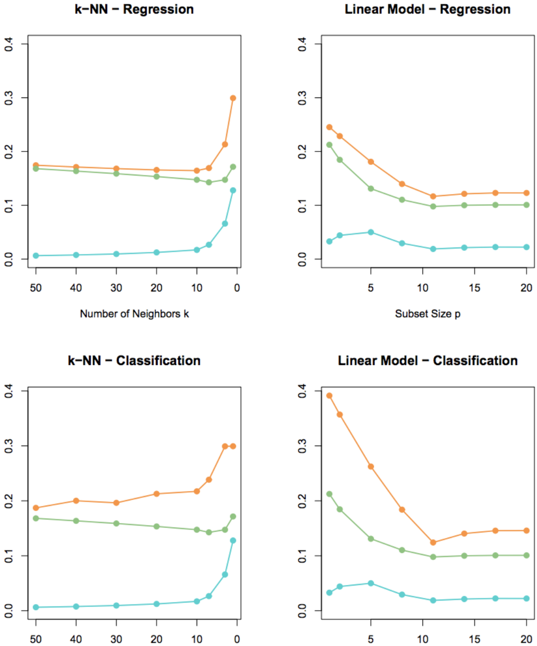
The bias variance tradeoff

- Suppose we choose to select our estimator based on a quadratic loss, $L(\hat{\theta}, \theta^*) = (\theta^* - \hat{\theta})^2$
- The corresponding function $\mathbb{E}\left\{(\theta^* - \hat{\theta})^2\right\}$ is known as the MSE

$$
\begin{aligned}
&\mathbb{E}\left\{\left[(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta^*)\right]^2\right\} \\
&\mathbb{E}\left[(\hat{\theta} - \bar{\theta})^2\right] + 2(\bar{\theta} - \theta^*)\mathbb{E}\left\{\hat{\theta} - \theta^*\right\} + (\bar{\theta} - \theta^*)^2 \\
&= \mathbb{E}\left\{(\hat{\theta} - \bar{\theta})\right\} + (\bar{\theta} - \theta^*)^2 \\
&= \text{var}(\hat{\theta}) + \text{bias}^2(\hat{\theta})
\end{aligned}
$$

- MSE $=$ variance $+$ bias^2

The bias variance tradeoff (H,T,F, Elem. of stat. Learning)



Empirical risk (I)

- In practice, we cannot compute the risk function in the frequentist setting
- It is however possible to optimize the loss in regression/prediction problems instead of hidden variables estimation
- In regression, we have a loss of the form $L(y, \delta_{\hat{\theta}}(\boldsymbol{x}))$ (we minimize mismatch between labels)

$$R(p_*, \delta) = \mathbb{E}_{x,y}[L(y, \delta_{\hat{\theta}})] = \sum_{\boldsymbol{x}} \sum_{y} L(y, \delta(\boldsymbol{x}))p^*(\boldsymbol{x}, y)$$

- $p_*(\boldsymbol{x}, y)$ is unknown but we can replace it by the empirical distribution

$$p^*(\boldsymbol{x}, y) = p_{\text{emp}}(\boldsymbol{x}, y|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{x}_i}\delta_{y_i}$$

Empirical risk (II)

- From the empirical distribution, we can define the empirical risk

$$R_{\text{emp}}(\mathcal{D}, \delta) = R(p_{\text{emp}}(\cdot|\mathcal{D}), \delta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \delta(\boldsymbol{x}_i))$$

- Taking $L(y, \delta(\boldsymbol{x})) = \mathbb{1}_{y \neq \delta(\boldsymbol{x})}$ gives the misclassification rate
- Taking $L(y, \delta(\boldsymbol{x})) = (y - \delta(\boldsymbol{x}))^2$ gives the mean sqaured error
- the optimal decision rule $\hat{\delta}$ is then obtained as

$$\hat{\delta}_{ERM} = \underset{\delta}{\text{argmin}} R_{\text{emp}}(\mathcal{D}, \delta)$$

Empirical risk (III)

- The empirical risk is equal to the Bayes risk if the prior on nature's distribution is that this distribution equals the empirical distribution.
- As a consequence, the empirical risk will lead to overfitting
- One solution is to add a regularizer (we will go back to this later when discussing regression)

$$R'(\mathcal{D}, \delta) = R_{\text{emp}}(\mathcal{D}, \delta) + \lambda C(\delta)$$

- $C(\delta)$ measures the complexity of the model.

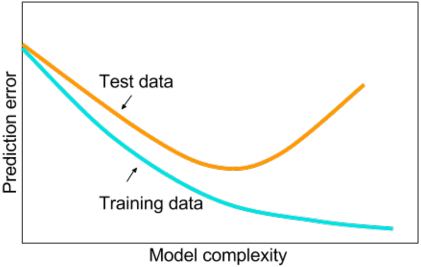
Empirical risk (III)

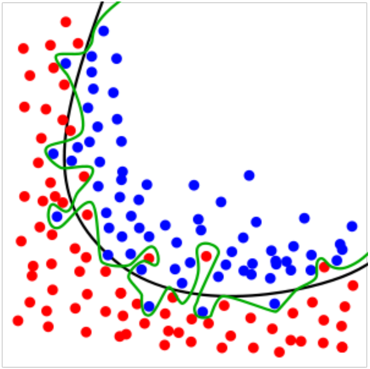
- The empirical risk is equal to the Bayes risk if the prior on nature's distribution is that this distribution equals the empirical distribution.
- As a consequence, the empirical risk will lead to overfitting
- One solution is to add a regularizer (we will go back to this later when discussing regression)

$$R'(\mathcal{D}, \delta) = R_{\mathsf{emp}}(\mathcal{D}, \delta) + \lambda C(\delta)$$

- $C(\delta)$ measures the complexity of the model.

Why is overfitting bad?





Structural risk and cross validation

- How do we choose the multiplier λ ?
- One possibility is to use cross validation.
  - Assume that you have a prediction model $\mathcal{P}(\boldsymbol{x}, \hat{\theta})$ which gives you outputs/labels

    $$\hat{y} = \mathcal{P}(\boldsymbol{x}, \hat{\theta})$$

  - Where $\hat{\theta}$ is estimated by fitting some a model of order (i.e complexity) m to the data

    $$\hat{\theta} = \mathcal{F}(\mathcal{D}, m)$$

  - We now split the dataset $\mathcal{D}$ into folds $\mathcal{D}_k$ and $\overline{\mathcal{D}}_k = \mathcal{D} \setminus \mathcal{D}_k$

Structural risk and cross validation

- From the folds $\mathcal{D}_k$, we can then write the risk R as

$$R(m, \mathcal{D}, K) = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} L(y_i, \mathcal{P}(\boldsymbol{x}_i, \mathcal{F}(\overline{\mathcal{D}}_k, m)))$$

- When taking $|\mathcal{D}_k| = 1$, this idea is called Leave one out cross validation
- In this case, the risk is simply given by

$$R(m, \mathcal{D}, N) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_m^{-i}(\boldsymbol{x}_i))$$

- How is that useful for regularization?

$$\hat{\lambda} = \operatorname*{argmin}_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}^k} L(y_i, f_\lambda^k(x_i))$$

Statistical Learning Theory

- Another approach would be to use statistical Learning theory (SLT) to derive an upper bound on the risk
- Will be covered in more details at the end of the class
- For today just remember that the deviation between the empirical risk and the population risk can be bounded as

$$P(\max_{h\in\mathcal{H}}|R_{\text{emp}(\mathcal{D},h)} - R(p^*, h)| > \varepsilon) \leq 2\text{dim}(\mathcal{H})e^{-2N\varepsilon^2}$$

Summary

Bayesian	**Frequentist**
Random parameters	Deterministic parameters but variability across trials
Optimizes posterior	optimizes loss
MAP	MLE
Max posterior is optimal	Consistency, variance, bias
Model selection through posterior expected loss	Model selection only possible in prediction through emp. loss