CS480/680: Introduction to Machine Learning
Lecture 04: Statistical Learning Basics

Yaoliang Yu

University of Waterloo

May 21, 2020

Distributions and density

The cumulative distribution function (cdf) of a random vector $\mathbf{X} \in \mathbb{R}^d$ is

$$F(\mathbf{x}) := \Pr(\mathbf{X} \leq \mathbf{x}),$$

and its probability density function (pdf) is

$$p(\mathbf{x}) := \frac{\partial^d F}{\partial x_1 \cdots \partial x_d}(\mathbf{x}), \text{ or equivalently } F(\mathbf{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} p(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

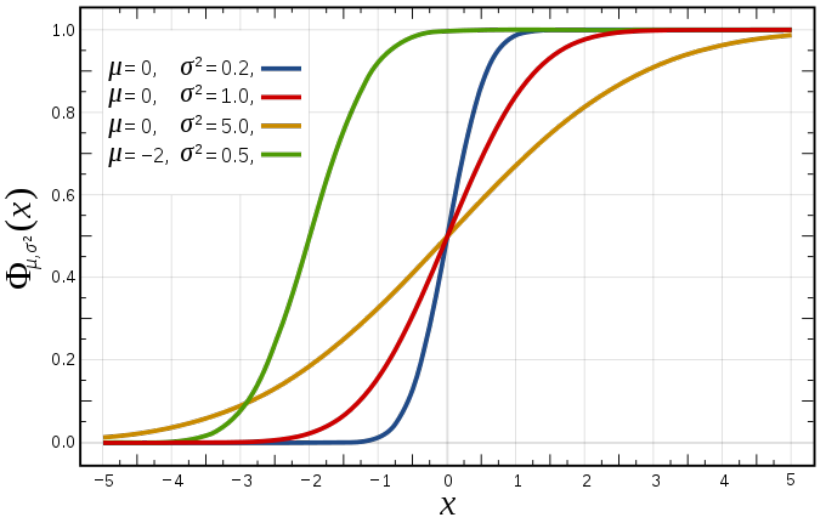
Clearly, each cdf $F : \mathbb{R}^d \to [0, 1]$ is

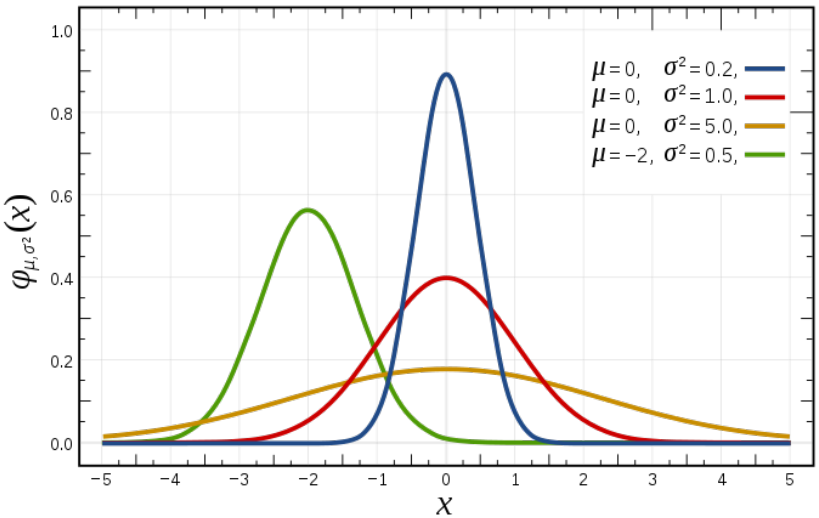
- monotonically increasing in each of its inputs;
- right continuous in each of its inputs;
- $\lim_{\mathbf{x} \to \infty} F(\mathbf{x}) = 1$ and $\lim_{\mathbf{x} \to -\infty} F(\mathbf{x}) = 0$.

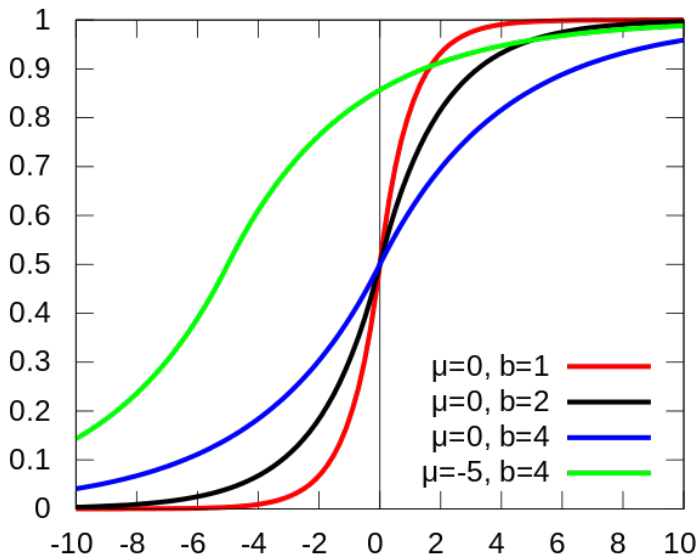
On the other hand, each pdf $p : \mathbb{R}^d \to \mathbb{R}_+$

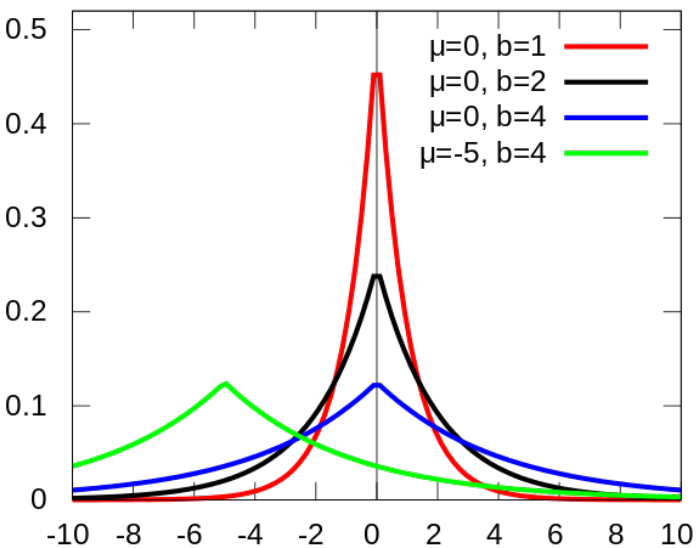
- integrates to 1, i.e. $\int_{-\infty}^{\infty} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 1$.

Univariate Gaussians and Laplacians









Change-of-variable

Let $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$ be a diffeomorphism (differentiable bijection with differentiable inverse). Let $\mathbf{X} = \mathsf{T}(\mathbf{Z})$, then we have the change-of-variable formula for the pdfs:

$$p(\mathbf{x})\,\mathrm{d}\mathbf{x} \approx q(\mathbf{z})\,\mathrm{d}\mathbf{z},\ i.e.\ p(\mathbf{x}) = q(\mathsf{T}^{-1}(\mathbf{x}))\left|\det \frac{\mathrm{d}\mathsf{T}^{-1}}{\mathrm{d}\mathbf{x}}(\mathbf{x})\right|$$

$$q(\mathbf{z}) = p(\mathsf{T}(\mathbf{z}))\left|\det \frac{\mathrm{d}\mathsf{T}}{\mathrm{d}\mathbf{z}}(\mathbf{z})\right|,$$

where $\det$ denotes the determinant.

Marginal and conditional

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ be a random vector with pdf $p(\mathbf{x}) = p(\mathbf{x}_1, \mathbf{x}_2)$.

- $\mathbf{X}_1$ is a marginal of $\mathbf{X}$ with pdf

$$p_1(\mathbf{x}_1) = \int_{-\infty}^{\infty} p(\mathbf{x}_1, \mathbf{x}_2)\, \mathrm{d}\mathbf{x}_2,$$

  where we marginalize over $\mathbf{X}_2$ by integrating it out.

- Define the conditional $\mathbf{X}_1|\mathbf{X}_2$ with density:

$$p_{1|2}(\mathbf{x}_1|\mathbf{x}_2) = p(\mathbf{x}_1, \mathbf{x}_2)/p_2(\mathbf{x}_2),$$

  where the value of $p_{1|2}$ is arbitrary if $p_2(\mathbf{x}_2) = 0$ (usually immaterial).

- Obvious from our definition that

$$p(\mathbf{x}_1, \mathbf{x}_2) = p_1(\mathbf{x}_1)p_{2|1}(\mathbf{x}_2|\mathbf{x}_1) = p_2(\mathbf{x}_2)p_{1|2}(\mathbf{x}_1|\mathbf{x}_2),$$

  namely the joint density p can be factorized into the product of marginal p_1 and conditional $p_{2|1}$.

Independence and chain rule

- Iterating the above construction, we obtain the famous chain rule:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d) = \prod_{j=1}^{d} p(\mathbf{x}_j | \mathbf{x}_1, \ldots, \mathbf{x}_{j-1}).$$

- We say that the random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d$ are independent if

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d) = \prod_{j=1}^{d} p(\mathbf{x}_j).$$

- The Bayes rule:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B, A) + \Pr(B, \neg A)}.$$

Mean, variance and covariance

Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random (column) vector.

We define its mean (vector) as

$$\boldsymbol{\mu} = \mathrm{E}\mathbf{X}, \quad \text{where} \quad \mu_j = \int x_j \cdot p(x_j)\,\mathrm{d}x_j$$

and its covariance (matrix) as

$$\Sigma = \mathrm{E}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top, \Sigma_{ij} = \int (x_i - \mu_i)(x_j - \mu_j) \cdot p(x_i, x_j)\,\mathrm{d}x_i\,\mathrm{d}x_j.$$

By definition Σ is symmetric $\Sigma_{ij} = \Sigma_{ji}$ and positive semidefinite (all eigenvalues are nonnegative).

The j-th diagonal entry of the covariance $\sigma_j^2 := \Sigma_{jj}$ is called the variance of X_j.

Multivariate Gaussian

The pdf of multivariate Gaussian/normal distribution $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$, with dimension d, mean $\boldsymbol{\mu}$ and covariance Σ, is:

$$p(\mathbf{x}) = (2\pi)^{-d/2}[\det(\Sigma)]^{-1/2} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Gaussians are equivariance under affine transformations:

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \implies A\mathbf{X} + \mathbf{b} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^\top).$$

(This property actually characterizes Gaussians.)

Let $\begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$.

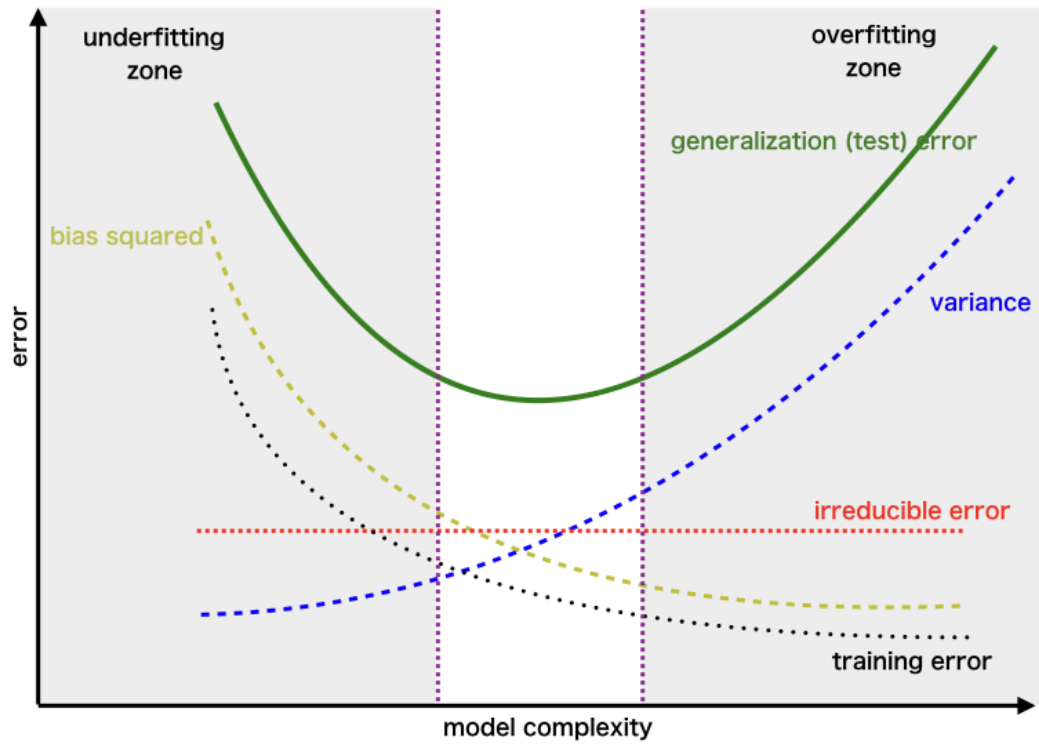
$$\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11}), \quad \mathbf{X}_2|\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12});$$
$$\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_{22}), \quad \mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Bias-variance trade-off

Predicting Y based on $\mathbf{X}$ under squared loss:

$$\mathsf{E}(\hat{f}(\mathbf{X}) - Y)^2 = \underbrace{\mathsf{E}\big(\hat{f}(\mathbf{X}) - \mathsf{E}\hat{f}(\mathbf{X})\big)^2}_{\text{variance}} + \underbrace{\mathsf{E}\big(\mathsf{E}\hat{f}(\mathbf{X}) - \mathsf{E}(Y|\mathbf{X})\big)^2}_{\text{bias}^2} + \underbrace{\mathsf{E}\big(\mathsf{E}(Y|\mathbf{X}) - Y\big)^2}_{\text{difficulty}},$$

where recall that $\mathsf{E}(Y|\mathbf{X})$ is the so-called regression function.



Maximum likelihood estimation (MLE)

Given $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $\mathbf{x}_i \overset{i.i.d.}{\sim} p(\mathbf{x}|\theta)$ with *unknown* parameter θ.

We define the likelihood of a parameter θ given the dataset $\mathcal{D}$ as:

$$L(\theta) = L(\theta; \mathcal{D}) := p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\theta).$$

A popular way to find an estimate of the parameter θ is to maximize the likelihood over some parameter space Θ:

$$\theta_{\text{MLE}} := \text{argmax}_{\theta \in \Theta} \; L(\theta).$$

Equivalently, by taking the log and negating, we minimize the negative log-likelihood (NLL):

$$\theta_{\text{MLE}} := \text{argmin}_{\theta \in \Theta} \; \sum_{i=1}^{n} -\log \, p(\mathbf{x}_i|\theta).$$

MLE is applicable only when we can evaluate the likelihood efficiently.

Sample mean and covariance as MLE

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

We apply maximum likelihood to estimate $\boldsymbol{\mu}$:

$$\hat{\boldsymbol{\mu}}_{\mathsf{MLE}} := \operatorname*{argmin}_{\boldsymbol{\mu}} \ \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}).$$

Applying Fermat's condition we obtain the sample mean:

$$\hat{\boldsymbol{\mu}}_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i =: \hat{\mathsf{E}}\mathbf{x}.$$

Similarly we can estimate Σ:

$$\begin{aligned} \hat{\Sigma}_{\mathsf{MLE}} &:= \operatorname*{argmin}_{\Sigma} \ \log\det\Sigma + \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &= \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top = \hat{\mathsf{E}}\mathbf{x}\mathbf{x}^\top - (\hat{\mathsf{E}}\mathbf{x})(\hat{\mathsf{E}}\mathbf{x})^\top, \end{aligned}$$

where we plug in the ML estimate $\hat{\boldsymbol{\mu}}_{\mathsf{MLE}}$ of $\boldsymbol{\mu}$ if it is not known.

f-divergence

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex function with $f(1) = 0$.
We define the f-divergence to measure the closeness of two pdfs p and q:

$$\mathsf{D}_f(p\|q) := \int f\big(p(\mathbf{x})/q(\mathbf{x})\big) \cdot q(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

where we assume $q(\mathbf{x}) = 0 \implies p(\mathbf{x}) = 0$ (otherwise we put the divergence to ∞).

- Let $f(t) = t \log t$, then we obtain the Kullback-Leibler (KL) divergence:

$$\mathsf{KL}(p\|q) = \int p(\mathbf{x}) \log(p(\mathbf{x})/q(\mathbf{x}))\,\mathrm{d}\mathbf{x}.$$

- On the other hand, let $f = -\log$ we obtain the reverse KL divergence:

$$\mathsf{LK}(p\|q) := \mathsf{KL}(q\|p).$$

Information theory

We define the entropy of a random vector $\mathbf{X}$ with pdf p as:

$$\mathsf{H}(\mathbf{X}) := \mathsf{E} - \log p(\mathbf{X}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) \, \mathrm{d}\mathbf{x},$$

the conditional entropy between $\mathbf{X}$ and $\mathbf{Z}$ (with pdf q) as:

$$\mathsf{H}(\mathbf{X}|\mathbf{Z}) := \mathsf{E} - \log p(\mathbf{X}|\mathbf{Z}) = -\int p(\mathbf{x}, \mathbf{z}) \log p(\mathbf{x}|\mathbf{z}) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{z},$$

and the cross-entropy between $\mathbf{X}$ and $\mathbf{Z}$ as:

$$\dagger(\mathbf{X}, \mathbf{Z}) := \mathsf{E} - \log q(\mathbf{X}) = -\int p(\mathbf{x}) \log q(\mathbf{x}) \, \mathrm{d}\mathbf{x}.$$

Finally, we define the mutual information between $\mathbf{X}$ and $\mathbf{Z}$ as:

$$\mathsf{I}(\mathbf{X}, \mathbf{Z}) := \mathsf{KL}(p(\mathbf{x}, \mathbf{z}) \| p(\mathbf{x}) q(\mathbf{z})) = \int p(\mathbf{x}, \mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x}) q(\mathbf{z})} \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{z}$$

MLE = KL minimization

Let us define the empirical "pdf" based on a dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$:

$$\hat{p}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i},$$

where $\delta_{\mathbf{x}}$ is the "illegal" delta mass concentrated at $\mathbf{x}$.

We claim that

$$\theta_{\mathsf{MLE}} = \operatorname*{argmin}_{\theta \in \Theta} \ \mathsf{KL}\big(\hat{p} \| p(\mathbf{x}|\theta)\big).$$

Indeed, we have

$$\mathsf{KL}(\hat{p} \| p(\mathbf{x}|\theta)) = \int [\log(\hat{p}(\mathbf{x})) - \log p(\mathbf{x}|\theta)] \hat{p}(\mathbf{x}) \, \mathrm{d}\mathbf{x} = C + \frac{1}{n} \sum_{i=1}^{n} - \log p(\mathbf{x}_i|\theta),$$

where C is a constant that does not depend on θ.

Linear regression as MLE

Let us now give linear regression a probabilistic interpretation. Assume:

$$Y = \mathbf{x}^\top \mathbf{w} + \epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Given a dataset $\mathcal{D} = \wr(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)\int$, the likelihood function of the parameter $\mathbf{w}$ is:

$$L(\mathbf{w}; \mathcal{D}) = p(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}\right)$$

$$\hat{\mathbf{w}}_{\mathsf{MLE}} = \operatorname*{argmin}_{\mathbf{w}} \quad \frac{n}{2} \log \sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2,$$

which is exactly linear regression (after ignoring irrelevant constants).

We can now also obtain an MLE of the noise variance σ^2:

$$\hat{\sigma}^2_{\mathsf{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\mathbf{w}}_{\mathsf{MLE}})^2,$$

which is nothing but the average training error.

Prior & Posterior

In a full Bayesian approach, we also assume the parameter θ is random and follows a prior pdf $p(\theta)$.

Ideally, we choose the prior $p(\theta)$ to encode our *a priori* knowledge of the problem at hand. (Regrettably, in practice computational convenience often dominates the choice of the prior.)

Suppose we have chosen a prior pdf $p(\theta)$ for our parameter of interest θ. After observing some data $\mathcal{D}$, our belief on the probable values of θ will have changed, so we obtain the posterior:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\,\mathrm{d}\theta},$$

where recall that $p(\mathcal{D}|\theta)$ is exactly the likelihood of θ given the data $\mathcal{D}$.

Computing the denominator (a.k.a. evidence) may be difficult since it involves an integral that may not be tractable.

Bayesian linear regression

Let us consider linear regression again:

$$Y = \mathbf{x}^\top \mathbf{w} + \epsilon, \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

This time we also impose a Gaussian prior on the weights $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbb{I})$. As usual we assume $\boldsymbol{\epsilon}$ is independent of $\mathbf{w}$.

Given a dataset $\mathcal{D} = \wr(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\int$, we compute the posterior:

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}) &\propto p(\mathbf{w})p(\mathcal{D}|\mathbf{w}) \\ &\propto \exp\left(-\frac{\lambda \mathbf{w}^\top \mathbf{w}}{2}\right) \cdot \prod_{i=1}^{n} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \mathbf{w})^2}{2\sigma^2}\right) \\ &= \mathcal{N}(\boldsymbol{\mu}_n, S_n), \end{aligned}$$

where (by completing the square) we have (with $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$)

$$S_n^{-1} = \lambda \mathbb{I} + X^\top X/\sigma^2$$

$$\boldsymbol{\mu}_n = S_n X^\top \mathbf{y}/\sigma^2.$$

We can also derive the predictive distribution on a new input $\mathbf{x}$.

Maximum a posteriori (MAP)

Another popular parameter estimation algorithm is the MAP that simply maximizes the posterior:

$$\begin{aligned}\theta_{\mathsf{MAP}} &:= \operatorname*{argmax}_{\theta \in \Theta} \; p(\theta|\mathcal{D}) \\ &= \operatorname*{argmin}_{\theta \in \Theta} \quad \underbrace{-\log p(\mathcal{D}|\theta)}_{\text{negative log-likelihood}} \quad + \quad \underbrace{-\log p(\theta)}_{\text{prior as regularization}}\end{aligned}$$

A strong (i.e. sharply concentrated, i.e. small variance) prior helps reducing the variance of our estimator, but potentially increasing our bias (cf. bias-variance trade-off) if our *a priori* belief is mis-specified.

Ridge regression as MAP

Let us consider linear regression one last time.

Like in Bayesian LR, impose $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbb{I})$. Then,

$$\hat{\mathbf{w}}_{\mathsf{MAP}} = \operatorname*{argmin}_{\mathbf{w}} \quad \frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 - \frac{d}{2}\log\lambda,$$

which is exactly equivalent to ridge regression.

Regularization constant λ is inverse to the variance of the prior. In other words, larger regularization means more determined prior information.

If we choose a different prior on the weights, such as the Laplacian prior, we obtain the celebrated Lasso:

$$\min_{\mathbf{w}} \quad \frac{1}{2\sigma^2}\|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1.$$

Bayes classifier

Consider the classification problem with random variables $\mathbf{X} \in \mathbb{R}^d$ and $Y \in [\mathsf{c}] := \{1, \ldots, \mathsf{c}\}$. The optimal (Bayes) classification rule is:

$$\begin{aligned}\hat{Y}(\mathbf{X}) &= \operatorname*{argmax}_{k \in [\mathsf{c}]} \Pr(Y = k | \mathbf{X}) \\ &= \operatorname*{argmax}_{k \in [\mathsf{c}]} \underbrace{\Pr(\mathbf{X} | Y = k)}_{\text{likelihood}} \cdot \underbrace{\Pr(Y = k)}_{\text{prior}},\end{aligned}$$

where ties can be broken arbitrarily.

In practice, we do not know the distribution of $(\mathbf{X}, Y)$, hence we cannot compute the optimal Bayes classification rule. One natural idea is to estimate the pdf of $(\mathbf{X}, Y)$ and then plug into above.

This approach however does not scale to high dimensions and we will see direct methods that avoid estimating the pdf.

Bayes rule arose from optimization

Let $p(\theta)$ be a prior pdf of our parameter θ, $p(\mathcal{D}|\theta)$ the pdf of data $\mathcal{D}$ given θ, and $p(\mathcal{D}) = \int p(\theta)p(\mathcal{D}|\theta)\,\mathrm{d}\theta$ the data pdf. Then,

$$p(\theta|\mathcal{D}) = \operatorname*{argmin}_{q(\theta)} \quad \mathsf{KL}\big(p(\mathcal{D})q(\theta) \parallel p(\theta)p(\mathcal{D}|\theta)\big),$$

where the minimization is over all pdf $q(\theta)$.

- If we restrict the minimization to a subclass $\mathcal{P}$ of pdfs, then we obtain some KL projection of the posterior $p(\theta|\mathcal{D})$ to the class $\mathcal{P}$. This is essentially the so-called variational inference.
- If we replace the KL divergence with any other f-divergence, the same result still holds. This opens a whole range of possibilities when we can only optimize over a subclass $\mathcal{P}$ of pdfs.
- The celebrated expectation-maximization (EM) algorithm also follows, as we will see!