A

Notation

The notation used in this work is very similar to the machine learning standard (for example, [20]). The subscript k always refers to the kth classifier, and the subscript n refers to the nth observation. The only exception is Chapter 5 that discusses a single classifier, which makes the use of k superfluous. Composite objects, like sets, vectors and matrices, are usually written in bold. Vectors are usually column vectors and are denoted by a lowercase symbol; matrices are denoted by an uppercase symbol. $\cdot^T$ is the transpose of a vector/matrix. $\hat{\cdot}$ is an estimate. $\cdot^*$ in Chapter 7 denotes the parameters of the variational posterior, and the posterior itself, and in Chapter 9 indicates optimality.

The tables in the next pages give the used symbol in the first column, a brief explanation of its meaning in the second column, and — where appropriate — the section number that is best to consult with respect to this symbol in the third column.

Sets, Functions and Distributions

$\emptyset$	empty set	
$\mathbb{R}$	set of real numbers	
$\mathbb{N}$	set of natural numbers	
$\mathbb{E}_X(X,Y)$	expectation of X, Y with respect to X	
$\mathrm{var}(X)$	variance of X	
$\mathrm{cov}(X,Y)$	covariance between X and Y	
$\mathrm{Tr}(\mathbf{A})$	trace of matrix $\mathbf{A}$	
$\langle \mathbf{x}, \mathbf{y} \rangle$	inner product of $\mathbf{x}$ and $\mathbf{y}$	5.2
$\langle \mathbf{x}, \mathbf{y} \rangle_A$	inner product of $\mathbf{x}$ and $\mathbf{y}$, weighted by matrix $\mathbf{A}$	5.2
$\\|\mathbf{x}\\|_A$	norm of $\mathbf{x}$ associated with inner product space $\langle \cdot, \cdot \rangle_A$	5.2
$\\|\mathbf{x}\\|$	Euclidean norm of $\mathbf{x}$, $\\|\mathbf{x}\\| \equiv \\|\mathbf{x}\\|_I$	5.2
$\\|\mathbf{x}\\|_\infty$	maximum norm of $\mathbf{x}$	9.2.1
$\otimes, \oslash$	multiplication and division operator for element-wise matrix and vector multiplication/division	8.1
L	loss function, $\mathrm{L} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$	3.1.1
l	log-likelihood function	4.1.2
$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$	4.2.1
$\mathrm{Gam}(x \mid a, b)$	gamma distribution with shape a, scale b	7.2.3
$\mathrm{St}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}, a)$	Student's t distribution with mean vector $\boldsymbol{\mu}$, precision matrix $\boldsymbol{\Lambda}$, and a degrees of freedom	7.4
$\mathrm{Dir}(\mathbf{x} \mid \boldsymbol{\alpha})$	Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$	7.5
p	probability mass/density	
q	variational probability mass/density	7.3.1
q^*	variational posterior	7.3
Γ	gamma function	7.2.3
ψ	digamma function	7.3.7
$\mathrm{KL}(q \\| p)$	Kullback-Leibler divergence between q and p	7.3.1
$\mathcal{L}(q)$	variational bound of q	7.3.1
$\mathbf{U}$	set of hidden variables	7.2.6

Data and Model

$\mathcal{X}$	input space	3.1
$\mathcal{Y}$	output space	3.1
$D_{\mathcal{X}}$	dimensionality of $\mathcal{X}$	3.1.2
$D_{\mathcal{Y}}$	dimensionality of $\mathcal{Y}$	3.1.2
N	number of observations	3.1
n	index referring to the nth observation	3.1
$\mathbf{X}$	set/matrix of inputs	3.1, 3.1.2
$\mathbf{Y}$	set/matrix of outputs	3.1, 3.1.2
$\mathbf{x}$	input, $\mathbf{x} \in \mathcal{X}$,	3.1
$\mathbf{y}$	output, $\mathbf{y} \in \mathcal{Y}$	3.1
$\boldsymbol{\upsilon}$	random variable for output $\mathbf{y}$	5.1.1
$\mathcal{D}$	data/training set, $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$	3.1
f	target function, mean of data-generating process, $f : \mathcal{X} \to \mathcal{Y}$	3.1.1
ϵ	zero-mean random variable, modelling stochasticity of data-generating process and measurement noise	3.1.1
$\mathcal{M}$	model structure, $\mathcal{M} = \{\mathbf{M}, K\}$	3.1.1, 3.2.5
$\boldsymbol{\theta}$	model parameters	3.2.1
$\hat{f}_{\mathcal{M}}$	hypothesis for data-generating process of model with structure $\mathcal{M}$, $\hat{f}_{\mathcal{M}} : \mathcal{X} \to \mathcal{Y}$	3.1.1
K	number of classifiers	3.2.2
k	index referring to classifier k	3.2.3

Classifier Model

$\mathcal{X}_k$	input space of classifier k, $\mathcal{X}_k \subseteq \mathcal{X}$	3.2.3
m_{nk}	binary matching random variable of classifier k for observation n	4.3.1
m_k	matching function of classifier k, $m_k : \mathcal{X} \to [0,1]$	3.2.3
$\mathcal{M}$	set of matching functions, $\mathcal{M} = \{m_k\}$	3.2.5
$\mathbf{M}_k$	matching matrix of classifier k	5.2.1
$\mathbf{M}$	matching matrix for all classifiers	8.1
$\boldsymbol{\theta}_k$	parameters of model of kth classifier	9.1.1
$\mathbf{w}_k$	weight vector of classifier k, $\mathbf{w}_k \in \mathbb{R}^{D_\mathcal{X}}$	4.2.1
$\boldsymbol{\omega}_k$	random vector for weight vector of classifier k	5.1.1
$\mathbf{W}_k$	weight matrix of classifier k, $\mathbf{W} \in \mathbb{R}^{D_\mathcal{Y} \times D_\mathcal{X}}$	7.2
τ_k	noise precision of classifier k, $\tau_k \in \mathbb{R}$	4.2.1
α_k	weight shrinkage prior	7.2
a_τ, b_τ	shape, scale parameters of prior on noise precision	7.2
a_{τ_k}, b_{τ_k}	shape, scale parameters of posterior on noise precision of classifier k	7.3.2
a_α, b_α	shape, scale parameters of hyperprior on weight shrinkage priors	7.2
$a_{\alpha_k}, b_{\alpha_k}$	shape, scale parameters of hyperposterior on weight shrinkage prior of classifier k	7.3.3
$\mathbf{W}$	set of weight matrices, $\mathbf{W} = \{\mathbf{W}_k\}$	7.2
$\boldsymbol{\tau}$	set of noise precisions, $\boldsymbol{\tau} = \{\tau_k\}$	7.2
$\boldsymbol{\alpha}$	set of weight shrinkage priors, $\boldsymbol{\alpha} = \{\alpha_k\}$	7.2
ϵ_k	zero-mean Gaussian noise for classifier k	5.1.1
c_k	match count of classifier k	5.2.2
$\boldsymbol{\Lambda}_k^{-1}$	input covariance matrix (for RLS, input correlation matrix) of classifier k	5.3.5
γ	step size for gradient-based algorithms	5.3
λ_{min} / λ_{max}	smallest / largest eigenvalue of input correlation matrix $c_k^{-1}\mathbf{X}^T\mathbf{M}_k\mathbf{X}$	5.3
T	time constant	5.3
λ	ridge complexity	5.3.5
λ	decay factor for recency-weighting	5.3.5
ζ	Kalman gain	5.3.6

Gating Network / Mixing Model

z_{nk}	binary latent variable, associating observation n to classifier k	4.1
r_{nk}	responsibility of classifier k for observation n, $r_{nk} = \mathbb{E}(z_{nk})$	4.1.3, 7.3.2
$\mathbf{v}_k$	gating/mixing vector, associated with classifier k, $\mathbf{v}_k \in \mathbb{R}^{D_V}$	4.1.2
β_k	mixing weight shrinkage prior, associated with classifier k	7.2
a_β, b_β	shape, scale parameters for hyperprior on mixing weight shrinkage priors	7.2
a_{β_k}, b_{β_k}	shape, scale parameters for hyperposterior on mixing weight shrinkage priors, associated with classifier k	7.3.5
$\mathbf{Z}$	set of latent variables, $\mathbf{Z} = \{z_{nk}\}$	4.1
$\mathbf{V}$	set/vector of gating/mixing vectors	4.1.2
$\boldsymbol{\beta}$	set of mixing weight shrinkage priors, $\boldsymbol{\beta} = \{\beta_k\}$	7.2
D_V	dimensionality of gating/mixing space	6.1
g_k	gating/mixing function (softmax function in Section 4.1.2, any mixing function in Chapter 6, otherwise generalised softmax function), $g_k : \mathcal{X} \to [0, 1]$	4.1.2, 4.3.1
ϕ	transfer function, $\phi : \mathcal{X} \to \mathbb{R}^{D_V}$	6.1
$\mathbf{\Phi}$	mixing feature matrix, $\mathbf{\Phi} \in \mathbb{R}^{N \times D_V}$	8.1
$\mathbf{H}$	Hessian matrix, $\mathbf{H} \in \mathbb{R}^{KD_V \times KD_V}$	6.1.1
E	error function of mixing model, $E : \mathbb{R}^{KD_V} \to \mathbb{R}$	6.1.1
γ_k	function returning quality metric for model of classifier k for state $\mathbf{x}$, $\gamma_k : \mathcal{X} \to \mathbb{R}^+$	6.2

Dynamic Programming and Reinforcement Learning

$\mathcal{X}$	set of states	9.1.1
$\mathbf{x}$	state, $\mathbf{x} \in \mathcal{X}$	9.1.1
N	number of states	9.1.1
$\mathcal{A}$	set of actions	9.1.1
a	action, $a \in \mathcal{A}$	9.1.1
$r_{xx'}(a)$	reward function, $r : \mathcal{X} \times \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$	9.1.1
$r^{\mu}_{xx'}$	reward function for policy μ	9.1.1
r^{μ}_{x}	reward function for expected rewards and policy μ	9.1.1
$\mathbf{r}^{\mu}$	reward vector of expected rewards for policy μ, $\mathbf{r}^{\mu} \in \mathbb{R}^N$	9.1.1
p^{μ}	transition function for policy μ	9.1.1
$\mathbf{P}^{\mu}$	transition matrix for policy μ, $\mathbf{P}^{\mu} \in [0,1]^{N \times N}$	9.1.4
γ	discount rate, $0 < \gamma \leq 1$	9.1.1
μ	policy, $\mu : \mathcal{X} \rightarrow \mathcal{A}$	9.1.1
V	value function, $V : \mathcal{X} \rightarrow \mathbb{R}$, V^* optimal, V^{μ} for policy μ, $\tilde{V}$ approximated	9.1.2
$\mathbf{V}$	value vector, $\mathbf{V} \in \mathbb{R}^N$, $\mathbf{V}^*$ optimal, $\mathbf{V}^{\mu}$ for policy μ, $\tilde{\mathbf{V}}$ approximated	9.1.4
$\tilde{\mathbf{V}}_k$	value vector approximated by classifier k	9.3.1
Q	action-value function, $Q : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$, Q^* optional, Q^{μ} for policy μ, $\tilde{Q}$ approximated	9.1.2
$\tilde{Q}_k$	action-value function approximated by classifier k	9.3.4
T	dynamic programming operator	9.2.1
T_{μ}	dynamic programming operator for policy μ	9.2.1
$\mathrm{T}^{(\lambda)}_{\mu}$	temporal-difference learning operator for policy μ	9.2.4
$\mathit{\Pi}$	approximation operator	9.2.3
$\mathbf{\Pi}_k$	approximation operator of classifier k	9.3.1
π	steady-state distribution of Markov chain $\mathbf{P}^{\mu}$	9.4.3
π_k	matching-augmented stead-state distribution for classifier k	9.4.3
$\mathbf{D}$	diagonal state sampling matrix	9.4.3
$\mathbf{D}_k$	matching-augmented diagonal state sampling matrix for classifier k	9.4.3
α	step-size for gradient-based incremental algorithms	9.2.6